AFRL-IF-RS-TR-2004-250
Final Technical Report
September 2004

RATS: REACTIVE ARCHITECTURES

USC Information Sciences Institute

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. J897

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

AFRL-IF-RS-TR-2004-250 has been reviewed and is approved for publication.

APPROVED: /s/
CHRISTOPHER J. FLYNN
Project Engineer

FOR THE DIRECTOR: /s/
JAMES A. COLLINS, Acting Chief
Information Technology Division
Information Directorate

DESTRUCTION NOTICE - For classified documents, follow the procedures in DOD 5200.22M. Industrial Security Manual or DOD 5200.1-R, Information Security Program Regulation. For unclassified limited documents, destroy by any method that will prevent disclosure of contents or reconstruction of the document.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2004	Final Mar 00 – Jul 03

4. TITLE AND SUBTITLE

RATS: REACTIVE ARCHITECTURES

5. FUNDING NUMBERS

C - F30602-00-1-0511
PE - 62301E
PR - JRAT
TA - S0
WU - 01

6. AUTHOR(S)

Marc Christensen, Fouad Kiamelev, Michael Haney, Charlie Kuznia and Stephen Crago

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

USC Information Sciences Institute
4676 Admiralty Way
Marina Del Rey, CA 90292-6695

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington, VA 22203-1714

AFRL/IFTC
26 Electronic Pky
Rome NY 13441-4514

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

AFRL-IF-RS-TR-2004-250

11. SUPPLEMENTARY NOTES

AFRL Project Engineer: Christopher Flynn, IFTC, 315-330-3249, flynnc@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

This project had two goals: To build an emulation prototype board for a tiled architecture and to demonstrate the utility of a global inter-chip free-space photonic interconnection fabric for polymorphous computer architectures (PCA). The free-space optics part of the project focused on two critical issues to validate the use of ultra-high-capacity global free-space interconnection fabrics for polymorphous computing architectures. These issues fell into the categories of interface issues and internal switch issues. The VCSEL-based free-space optical system was shown to be tolerant of misalignments, although electrical packaging problems prevented a full-performance demonstration. The optical links for the free-space system were demonstrated at 2.5 Gbps each with an inter-chip distance of about 10 cm. The project successfully demonstrated an optical inter-board fiber-based interconnect and intra-board free-space interconnect. The project also developed a prototype board that was demonstrated with an emulator of a tiled architecture.

14. SUBJECT TERMS

Photonics, free-space optical interconnect, polymorphous computing architectures

15. NUMBER OF PAGES

62

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

1.A Project Goals

The Reactive ArchiTectureS (RATS) project started as a seedling effort to explore reactive architectures, which came to be known as polymorphous computing architectures under the Polymorphous Computer Architectures (PCA) project sponsored by the Defense Advanced Research Project Agency (DARPA). Early in the project, the RATS team identified two important topics of research relevant to polymorphous computing. The first topic was tiled architectures, and the second was global interconnects that enable applications with global and/or dynamic communication requirements. In the area of tiled architectures, the RATS project collaborated with the Raw project at the Massachusetts Institute of Technology (MIT). MIT developed a programmable logic-based emulation of their Raw architecture, and the RATS project developed the Raw emulation board to allow functional testing of the Raw architecture. The RATS project also decided to demonstrate free-space optical interconnect technology to address the communication challenges inherent in polymorphous computing.

The goal of the free-space interconnect demonstration of the RATS project was to demonstrate the utility of a global inter-chip free-space photonic interconnection fabric for polymorphous computer architectures. The reactive nature of PCA computing dictates that a flexible interconnection fabric is highly desirable. For instance, as a collection of processors "morphs" in response to evolving application requirements the physical location of the processors in the system may become fragmented with respect to the logical location of the processors in the data flow graph, i.e. processors wishing to share data closely (logically near to each other) may be physically located far apart such that data transfers among them may be difficult or costly. Adding an interconnection fabric which can "morph" to meet the needs of an evolving computer architecture would bring a new dimension to the possible PCA solution space. Photonic interconnection fabrics are one technology which have the capacity to consider embedding a fully interconnected data flow graph, thereby enabling "morphing" among many different interconnection possibilities. While the large interconnection capacity of the photonic fabric is never directly utilized, it is required by the varied nature of possible desired configurations that must all be implemented by a single technology solution. To that end the research project described herein was conceived to show the readiness of photonic interconnection fabrics to be applied to future computer architecture problems. As these computer architectures will likely take a variety of form factors, the program was configured to address both systems that were embedded onto a single processing board and systems that required the connection of many such boards.

This program had two main objectives dealing with: 1) the critical interface to a massive global free-space interconnection fabric, and 2) validation of the global fabric itself. The first objective centered on showing the feasibility of a multi-board parallel fiber optical interconnection module for reactive processing with a specific goal of transmitting multi-gigabit data between a standard FPGA (Field Programmable Gate Array) board and a daughtercard module and across the parallel fiber optic link. The second specific objective centered on the demonstration of a large capacity (>100 Gigabits/second) free-space interconnection fabric for multiple processors on a single board.

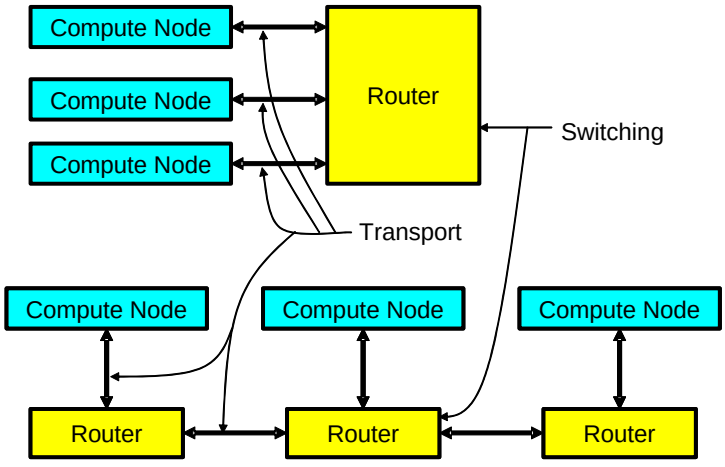


Figure 1-1. Interconnect terminology

I.B. Approach

We have developed a simple taxonomy that we will use to explore the space of inter-module interconnects. The taxonomy classifies interconnect designs according to how they implement two functions: data transport and data switching. Data transport is defined as the movement of data between compute nodes and routers or between routers, as shown in Figure 1-1. Data switching is defined as the routing of data from a source port to a destination port. For each interconnect function, there are two viable options for high performance/productivity computing systems: electrical and optical. Other technologies that provide these functions exist, for example proximity I/O and wireless communication, but they are not viable for high-speed inter-module interconnects. Other important interconnect variables exist, such as the number of parallel links per connection and wavelength (for optical communication), but they are not included in this top-level taxonomy. Table 1-1 shows our interconnect taxonomy. All-electrical interconnects are in the upper left quadrant of the taxonomy.

Table 1-1. Interconnect taxonomy

		Switching	
		Electrical	Optical
Transport	Electrical	Traditional electrical interconnects	Optical router w/ electrical I/O
	Optical	Free-space optical interconnects	Optical router w/ optical I/O

Most of today's high-performance systems are in the upper left quadrant. In the upper right corner are interconnects with electrical transport and optical switching. For reasons that will be explained below, interconnects in the upper right quadrant are not interesting. The lower left quadrant contains interconnects that optically transport data and electrically switch. Free-space optical interconnects and fiber-to-the-processor (which use optical signals to communicate between CMOS chips that implement computing, memory, and switching) interconnects fit into this quadrant. Finally, the quadrant in the lower right contains all optical interconnects. All optical interconnects are currently being developed for telecom markets.

Electrical and optical technologies each have their own strengths and these strengths have different effects in different quadrants of the taxonomy. Generally, electronic technologies are better than optical for implementing logic because electrons have inherent interactions that are useful for logical functions. Optical technologies have advantages for communication. Light can be transported over long distances with minimal signal degradation through free space or on fiberoptic cables. Photons do not generally interact with each other (ignoring quantum effects), so effects such as crosstalk are minimal. These generalities lead to the conclusion that switching should be done electrically and that transport should be done optically. However, second-order effects make these trade-offs more interesting. For example, data are generated electrically within a compute chip. There are costs to converting electrical signals to optical signals for transport. Also, if data are transported optically, there is a cost to converting it back to electricity to switch electrically.

Electrical transport, electrical switching

All-electrical interconnects are the most common interconnects in embedded systems, commodity workstations, and high-performance computing today. All electrical interconnects have several advantages. First, electrical interconnects have history and momentum on their side. Since computation is almost always done electrically (except for a few special-purpose optical computing applications), no new technologies are needed to drive signals electrically. When processing technology was much slower, driving electrical signals between chips or even modules at speeds close to processing speed was not a problem. However, as computing density has followed Moore's Law, two things have happened to make alternatives to all-electrical interconnects attractive. First, the number of pins per chip has not scaled with Moore's Law. Second, it has become much more difficult to electrically drive pins and connections at the frequencies of logic components. These problems have made optical interconnect technologies worth exploring, at least for inter-module communication. The communication distance for which optical communication is useful has been decreasing as frequencies increase and optical technology develops.

Electrical transport, optical switching

Interconnects with electrical transport and optical switching are the least interesting quadrant in the taxonomy. This combination uses each technology where it is weakest. Signals would be sent long distances using electrical technology that is hampered by transmission line effects and noise; and power and design complexity are required to overcome these effects. Then signals must be converted from electrical to optical, where photons must be “steered” in order to route them from source to destination.

Optical transport, optical switching

All-optical networks are attractive because they allow data to be transported optically and then switched without optical/electrical conversions. Data serialization/deserialization and other overheads at the switch are eliminated and the switching infrastructure runs at optical rates and potentially has bit-rate independent data paths. However, current technology does not allow for fast switching of all-optical networks, which eliminates them from consideration for multiprocessor interconnects of the kind being studied for the RATS project.

Optical transport, electrical switching

Interconnects based on optical transport with electrical switching leverage the strength of optics for transporting data and the strength of electronics for implementing the logic of switching. Schemes that fall into this quadrant include many free-space optical interconnects and interconnects that use fiber optics to transport data between CMOS switches. FAST-Net (Free-space Accelerator for Switching Terabit-NETworks) is an example of a free-space optical interconnect that is in the optical transport, electrical switching quadrant [30]. FAST-Net uses VCSEL-based (Vertical Cavity Surface Emitting Laser) free-space optics to send data from source to destination. However, the actual switching is done electronically at the source in a smart-pixel array when data are routed to a specific set of VCSELs based on the destination. Switching is not implemented optically since each VCSEL sends data to only one destination and once data are in optical form, their destination is pre-determined. Interconnects that use fiber optics between electronic switches are a natural extension to all-electrical networks. VCSEL-based parallel optics technology can replace long electrical connections to improve bandwidth, power, and noise characteristics. This is the approach that was used in the RATS project.

To illustrate the performance limitations of the electrical interconnect approach, we conducted a study of mesh-based electrical interconnects on the corner turn application kernel [1]. The corner turn is a distributed matrix transpose operation that is used in 2D-FFT computations and other applications where data is operated along multiple dimensions. The goal of a scalable computer architecture is to allow the size of an application (measured in operations or data size) to be scaled linearly with the number of compute nodes in a system. Mesh-based interconnects have the advantage that the interconnect link bandwidth scales linearly with the number of compute nodes. However, for a kernel such as corner turn that requires global communication that scales linearly with the data size, the network traffic required scales faster than the number of compute nodes and interconnect links. This is because the number of messages scales at the same rate as the number of compute nodes, but the average message requires more hops to go from the source node to the destination node. Figure I-2 shows that when a mesh gets as large as 16x16 nodes, the network will start to saturate and for larger meshes, the mesh cannot sustain the traffic required for a high-performance corner turn even under the optimistic assumption that the network runs at the same speed as the processor logic. The problem illustrated in Figure 1-2 can be addressed by implementing a network with a higher degree of connectivity. Since components must typically be laid out in a 2-D or 3-D array, higher degrees of connectivity require longer distances between adjacent nodes, which make optical transport more attractive.

- Links (architecture) = O(P)
- Number of messages (problem size scales w/ P) = O(P)
- Number of hops per message = O(SQRT(P))
- Traffic generated / Links = O(P*SQRT(P)/P) = O(SQRT(P))
- Performance limited by global network, not tile bandwidth or chip I/O

Mesh Size	Chips	Links	Avg. Hops	Total Traffic	Utilization
4x4	1	96	2.5	20	0.21
8x8	4	448	5.25	168	0.375
16x16	**16**	**1920**	**10.6**	**1357**	**0.71**
32x32	64	7936	21.3	10906	1.375
64x64	256	32256	42.7	87450	2.71

Figure 1-2. Corner turn on mesh utilization analysis

The combination of free-space optical interconnections (FSOI) with smart pixel technology (based on the integration of Silicon ICs with arrays of vertical cavity surface emitting lasers (VCSELs) and photodetectors is projected to enable chip-to-chip interconnection fabrics that achieve bandwidth densities on the order of a Terabits/sec/cm^2 [27]. Scaleable global (i.e., chips-to-chips) interconnection fabrics that achieve minimum bisection bandwidths in the multi-terabits/sec regime may be implemented using multiple optoelectronic integrated circuits linked to each other in the manner depicted in Figure 3-1 [28]. This approach is the basis for a global chips-to-chips interconnection approach termed FAST-Net (Free-space Accelerator for Switching Terabit Networks) [29, 30]. In this approach the optical I/O from any single smart pixel array (SPA) chip, located at a lens' focal plane, is linked to portions of the I/O arrays of all chips in the system. To achieve this, clusters of VCSELs and photodetectors are imaged onto corresponding clusters on other chips. Multiple point-to-point links are established between cluster pairs on different SPAs. The clusters are interleaved to achieve a global interconnection pattern across the multi-chip plane, thus implementing a high-density bi-directional data path between every pair of SPA chips on the module.

1.C. Summary of Accomplishments

The major research highlights and accomplishments of the RATS project are listed below:

- Development of RATS emulation board
- Development of a 10 Gbits/second Optical Network Interface Card demonstrating direct leveraging of photonic interconnection fabrics for multi-board systems

- Development of a multi-scale optical lens system that removes distortion for a free-space optical interconnection module
- Demonstration of a free-space optical interconnection assembly with misalignment sensitivities better than 25 microns and 1 degree
- Integration of 4 1x4 VCSEL arrays and 4 1x4 photodetector arrays on a single Ultra-thin Silicon on Sapphire (UTSi) circuit for an aggregate free-space I/O bandwidth of 40 Gbps per chip
- Extension of chip-on-pin concept for optoelectronic device packaging for fiber modules and ~10 cm diameter multi-chip arrays on a board
- Demonstration of 160 Gpbs free-space interconnected module

1.D. Report Organization

As described in Section I.A, the goals of this project were spread over a range of system aspects and configurations spanning from optical network interface cards for multi-board or multi-chassis configurations to validating key aspects of global inter-chip free-space optical interconnection fabrics. As such this final report is organized to show results obtained over the course of this program in each area and how the techniques can be combined in a high-performance system. Section II describes the design of a novel multi-gigabit optical network interface card that would be needed at the I/O ports to fully exploit the free-space optical switch core when embedded in PCA-like architectures in which computing resources are distributed in a multi-board configuration. Section 3 describes research done in optimizing the robustness of the RATS free-space optical interconnection scheme under packaging misalignments. Both global inter-chip fabrics and board-to-board high-bandwidth link architectures are considered. Section 4 describes experimental results from the final free-space demonstration module in which parallel 2.5-Gbps channels were demonstrated in various inter-chip global combinations. Sections V and VI describe low power protocol driver research that is needed to exploit the ultimate ultra-high-density free-space optical interconnects. Section VII describes the Raw emulation board developed by the RATS project in collaboration with MIT. Section VIII concludes the report by summarizing the accomplishments of this project.

2. DESIGN OF A MULTI-GIGABIT OPTICAL NETWORK INTERFACE CARD

2.A. Network Interface Card Overview

Before building a complete free-space optical system, the RATS team decided to build a demonstration system that showed the capabilities of commercial photonics technology. We decided to build a demonstration board that could be used as an interface in a computer to fiber-based optical network that supported flexibility in the network protocol.

Various very short-reach (VSR) optical data links that operate at data rates of 10Gb/s and beyond are now becoming available as commercial products [2][3][4]. Various network, protocol and switch architectures that utilize these links have been proposed [5][6][7]. An example network architecture, shown in Figure II-1, uses VSR optical data links to interconnect multiple compute nodes through a central switch. In order to efficiently utilize the increased bandwidth capability of VSR optical data links, these networks architectures use new communication protocols rather than relying on existing standards such as ATM (Asynchronous Transfer Mode), Ethernet, or HIPPI (HIgh Performance Parallel Interface) [8][9][10]. The optical network interface card (ONIC) is an important instrument for demonstrating efficient application of these new architectures. The purpose of the ONIC is to interface a standard computing node, such as a workstation or an embedded processor, with a VSR optical data link. On the hardware front, the ONIC converts slow, wide-parallel and clock-synchronous electrical data streams (typically used in chip-level computer interconnection) to narrow gigabit-speed optical data streams with embedded clock information (typically used in VSR optical data links). The ONIC hardware often contains First-In First-Out (FIFO) memory storage to buffer incoming and outgoing network data for flow control. On the software front, the ONIC includes software drivers to provide communication and flow control between the hardware, application running on the computing node and the custom network protocol used by the network that is being demonstrated.

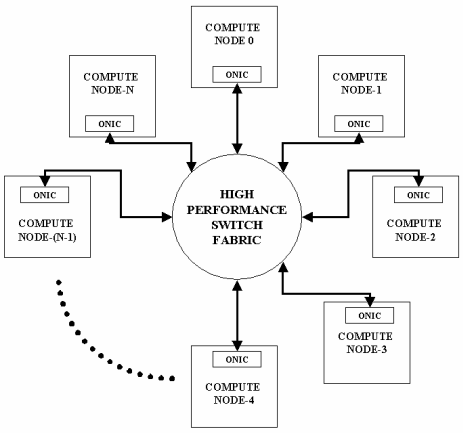


Figure 2-1. Switching system with its compute nodes communicating through ONIC

In this section, we describe the design of a programmable ONIC that interfaces a 12-channel, gigabit parallel optical link module [11][12] with a 64-bit/66-MHz PCI bus [13]. The adoption of the PCI bus allows our ONIC design to be used in widely available PCI-based workstation and server computers. Hardware programmability is achieved using field programmable gate array (FPGA) integrated circuits. This enables our ONIC design to efficiently implement different network protocols. Although the current ONIC design uses specific VSR optical data link hardware, it can be readily modified to support other types of optical data links. Our ONIC design was originally developed to demonstrate a specific network architecture that used free-space optical interconnection inside the switch fabric [5]. However, the novelty of our design is that it provides a low-cost network interface solution that can be readily modified by other researchers for network protocol and optical device specific requirements. The following paragraphs compare our approach with existing ONIC implementations and provide justification for our approach.

Commercially available ONICs use custom integrated circuits and/or network processor chips to implement specific network protocols [14][15]. They cannot keep up with the 10-100 Gigabit data rates available with VSR optical data links. Finally, commercial ONICs use proprietary designs and cannot be modified to use new VSR optical link hardware. These attributes make commercial ONICs unusable for experimenting with new network architectures and VSR optical data links. On the other hand, an ONIC has been previously demonstrated that used custom-made integrated circuits to implement a specific network protocol [16]. While this approach successfully demonstrated the new network protocol proposed by its authors, it is difficult to modify this design because of the high cost and extensive knowledge required for the design of custom integrated circuits.

References [6] and [17] describe a network architecture demonstration that used an ONIC design similar to the one being proposed in this section. That ONIC design also used FPGAs and SERDES (SERializer/DESerializer) that can be reconfigured to support various network protocols. However, it employed a proprietary memory-bus interface to connect the optical link with the computing node. While a memory bus interface permits higher bandwidth communication between the processor and the optical link, our PCI-based ONIC supports a broad range of computing hardware from a multitude of

computer manufacturers. Future modifications of our ONIC design can use the emerging PCI bus extensions [18] to achieve higher communication bandwidth than that possible with existing 64-bit/66-MHz PCI standard. Finally, we have done more work at the software driver level to demonstrate application-level message passing between computing nodes interconnected using ONICs.

The remainder of this section is organized as follows: Section 2.B describes the ONIC architecture and design. Section 3.C describes the experimental tests performed with prototype ONICs and the results. Section 2.D describes the use of ONIC in a new network architecture. Finally, Section 2.E provides concluding remarks.

2.B. ONIC Architecture and Design

The ONIC architecture is shown in Figure 2-2. It is a PCI based Network Interface Card (NIC). It has a high-end FPGA that can be programmed with the network protocols. The network interface of the ONIC is through two 12-channel gigabit parallel optical link modules to transmit and receive data. The FPGA sends and receives the parallel data from the compute server to the optical modules through a set of SERDES (SERialize-DESerialize) modules. The SERDES modules convert the parallel data to high-speed serial data suitable for the optical transmit module and recover the data and clock from the optical receive module. The following paragraphs describe the functionalities of the chips on the ONIC, hardware programming steps and physical layout of the Printed Circuit Board (PCB).

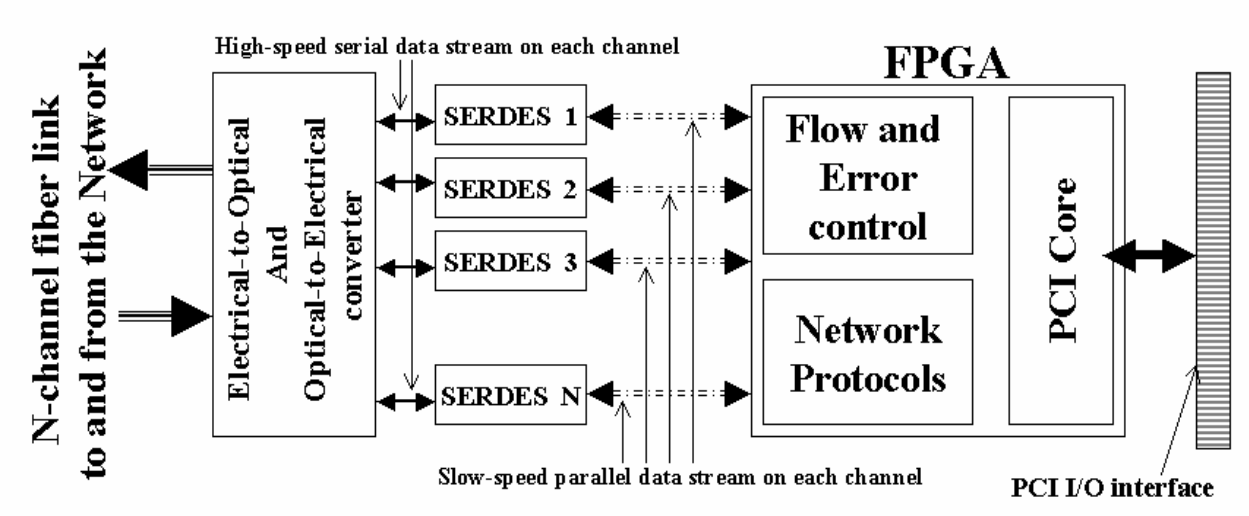


Figure 2-2. Architecture of the ONIC

The schematic of the ONIC is shown in Figure 3-3. The communication between the ONIC and the compute server is through the fast-wide PCI 64-bit/66MHz interface. The theoretical maximum bandwidth of the bi-directional PCI interface is 4.22Gb/s half-duplex. The FPGA is a high-performance VIRTEX XCV1000 from Xilinx having a capacity of more than one million system gates. The whole network protocol can be implemented on the FPGA, which gives greater flexibility. Some of the FPGA resources on the ONIC were devoted to Xilinx's PCI core.

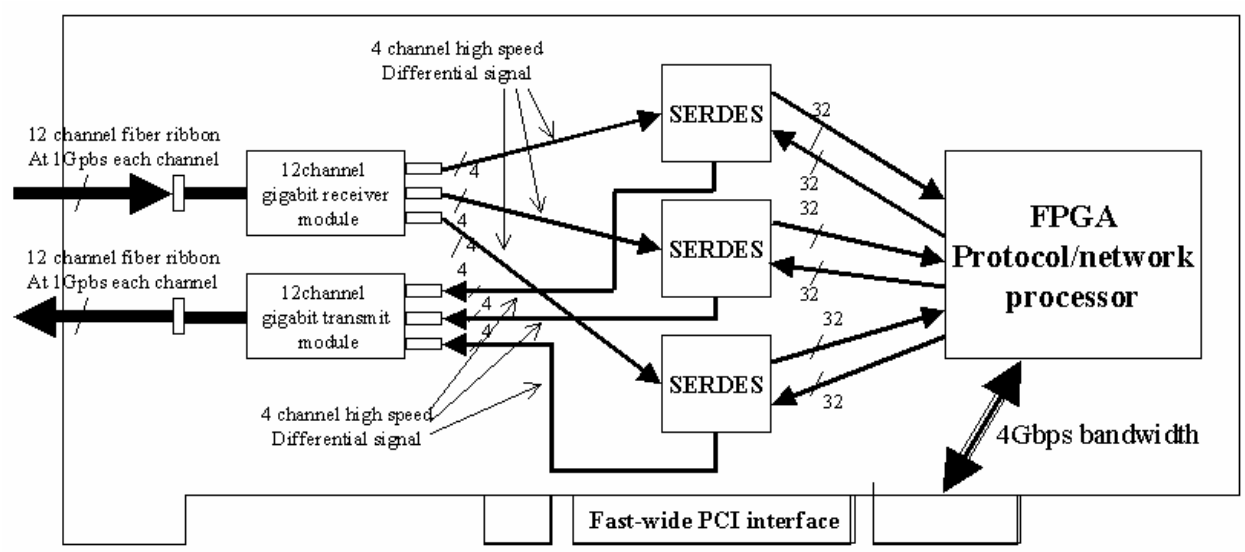


Figure 2-3. Schematic of the ONIC prototype hardware

The data from the FPGA goes to three AMCC S2065 quad-channel serial back-plane SERDES. Each SERDES has a 32-bit slow-speed input and output channel and four high-speed differential inputs and outputs that can operate between 0.7Gb/s and 1.3Gb/s. The data to and from the compute server to the SERDES is from the FPGA through the 32 data lines running between 70Mb/s and 130Mb/s. The SERDES on the transmit side does an 8b/10b encoding of the data. It generates K28.5 synchronization characters to establish communication with the destination node.

The high-speed differential signals from the SERDES connect to a 12-channel gigabit parallel optical link driver and receiver module from Honeywell Technology Center. Each module couples to a 12-channel parallel fiber link. Each channel of this link has been demonstrated at 1.06 Gb/s giving the composite parallel link a full-duplex bandwidth of 24 Gb/s. With the 8b/10b encoding/decoding of the data done by the SERDES at the transmitting/receiving ONICs to maintain signal integrity, the real full-duplex data bandwidth is 24 Gb/s for each compute server. The data from the optical receive module is sent to the SERDES. The SERDES does data decoding, clock recovery and synchronization.

The FPGA on the ONIC is programmed using a standard VHDL-based (Very high-speed integrated circuits Hardware Description Language) design flow. A JTAG (Joint Test Action Group) interface is provided for PROM programming. The ONIC was fabricated using a standard copper and FR-4 PCB fabrication process. The printed circuit board (PCB) has eight layers of routing with full and split power planes. Figure 2-4 shows a picture of the fabricated ONIC hardware. The ONIC draws more power than the PCI bus can sustain. The SERDES draws close to 10 watts of power at peak performance. Hence the ONIC is powered externally from the standard PC power supply. This isolation of power from PCI bus ensures reliable power without affecting the ability to add other system components on the PCI bus. On board DC/DC converters are used to regulate the power to various components on the ONIC.

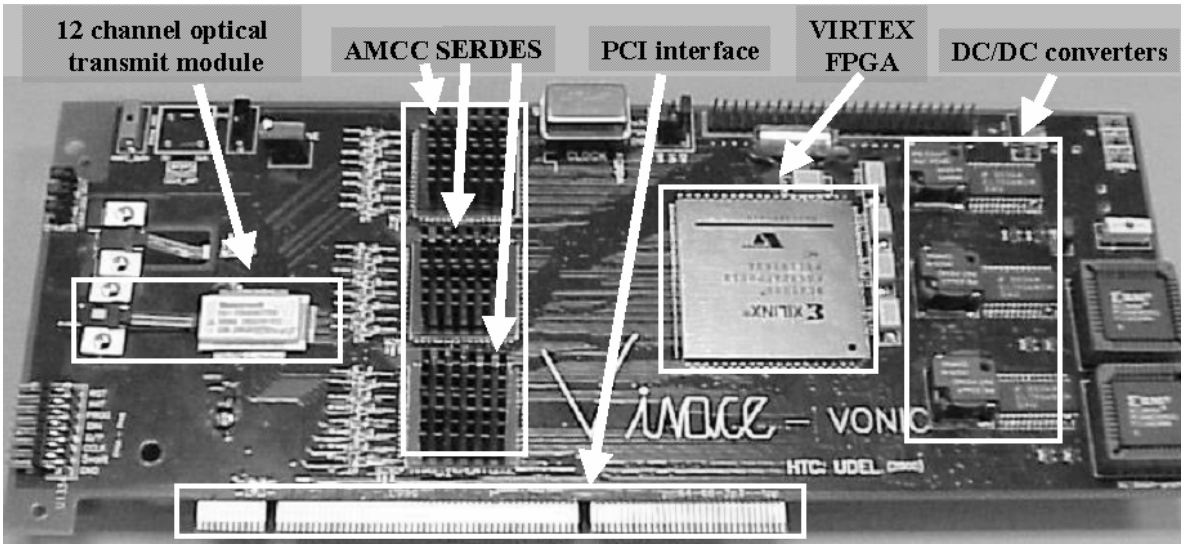


Figure 2-4. Fabricated ONIC prototype hardware

The parallel optical driver and receiver modules are located on either side of the ONIC. The Transmit module is shown in Figure 2-4. The receiver module is attached on the other side of the ONIC and the fiber link comes through a cut out on the card. Differential transmission lines run from the SERDES to the optical transmit and receive modules. These lines are 50ohm impedance matched and they are located only on the outer layer of the ONIC. This was done to avoid multiple vias. Figure 2-5 is a snapshot of the high-speed traces on the ONIC.

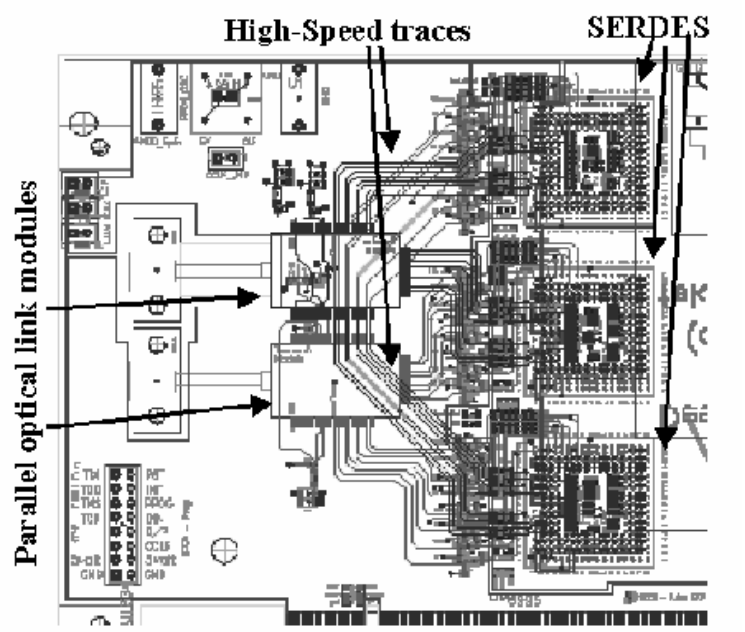


Figure 2-4. Impedance matched high-speed traces on the ONIC

2.C. Test Results

The ONIC hardware was tested in two phases. A link integrity test for Bit Error Rate (BER) analysis and a message passing application were run to test the ONIC hardware. These tests were conducted by sending synchronization characters followed by a digital signature (data starting point). This is followed by the real data.

2.C.1. Case I: Link Integrity Test

The link integrity test was performed on the ONIC hardware for BER analysis. Initially, the ONIC loop-back test was performed. Only one ONIC was used in this test. The parallel fiber ribbon from the transmitting optical module is looped back to the optical receiver module. The fiber ribbon was one meter in length. The ONIC was plugged into the PCI bus of a compute server. The FPGA of the ONIC was programmed with the test protocols and the PCI core. A 32-bit Linear Feedback Shift Register (LFSR) was used to generate Pseudo Random Bit Stream (PRBS).

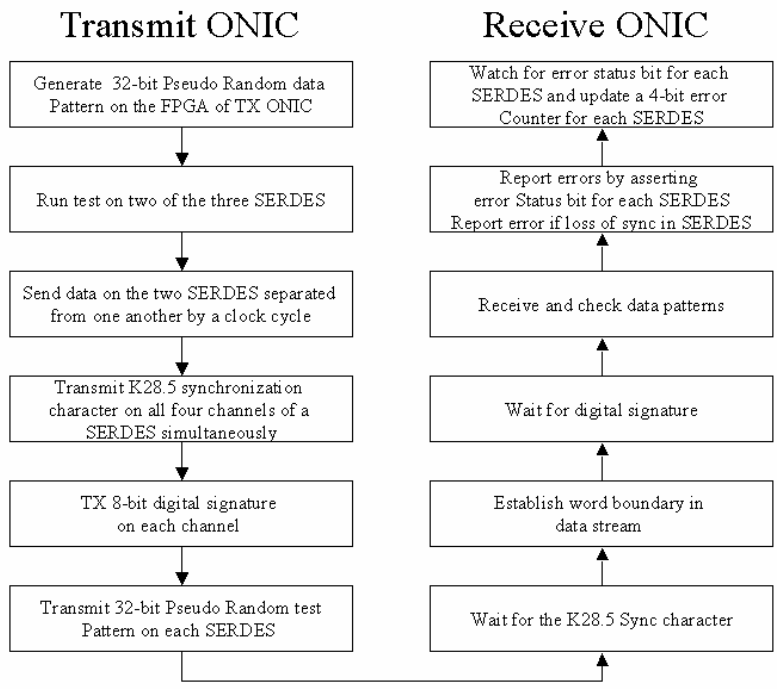


Figure 2-5. Link integrity test sequence

All three SERDES on the ONIC are controlled by the FPGA. Two of the three SERDES were run in sync with each other. The third SERDES was not used in the test. The K28.5 character for data synchronization between the two ONICs was sent on all four channels of each SERDES. After sending the K28.5 characters, an 8-bit digital signature was sent. This was followed by the same PRBS on both SERDES delayed from one another by a clock cycle. This data is encoded by the SERDES and sent to the optical modules. The data is then sent through the fiber-ribbon and is received by the receiver module. Figure 2-6 shows the steps involved in the link integrity test.

The data goes to the receiver optical module and is converted to electrical signals and sent to the SERDES. The SERDES waits for the K28.5 characters and when it recognizes the characters it synchronizes with the oncoming data. The SERDES performs data decoding, recovers the clock from the data, and sends them to the FPGA. The FPGA establishes the word boundary in the data stream from the K28.5 characters and waits for the digital signature. As soon as the FPGA sees the digital signature, it starts an LFSR with the same seed as the transmit LFSR and compares the oncoming data with its LFSR data. A status bit gets asserted when an error occurs. An error counter keeps track of the errors from each SERDES. The FPGA also reports an error when synchronization is lost in any of the SERDES. Figure 2-7 shows a scope snapshot of the data through one of the high-speed channels at 1 Gb/s. The error count is communicated back to the compute server through the PCI interface and is continuously updated. The test was run for 30 days and no errors were encountered. The total bandwidth of optical data communication in this test was 8Gb/s. The same test was repeated with two ONICs plugged onto two compute servers. One ONIC acts as the transmitter and the second ONIC as the receiver. The protocols were separated and the transmitting ONIC was programmed with the PCI core and the transmission protocol. The receiver ONIC was programmed with the PCI core and the receiver protocol. The length of the fiber ribbon used in this test was 1 meter. The test was run for 2 hours without any errors.

2.C.2. Case II: Message Passing Application

A message passing demonstration was performed to test application level communication between two ONICs plugged into the PCI bus of two compute servers. They were connected to each other through a 12-channel fiber-link. Figure 2-8 shows a picture of the two servers with ONIC hardware. A custom software driver was written for the application to communicate with the PCI core on the ONIC FPGA through the PCI bus. The device driver was developed for Windows NT 4.0. The server uses the driver function calls to communicate with the ONIC.

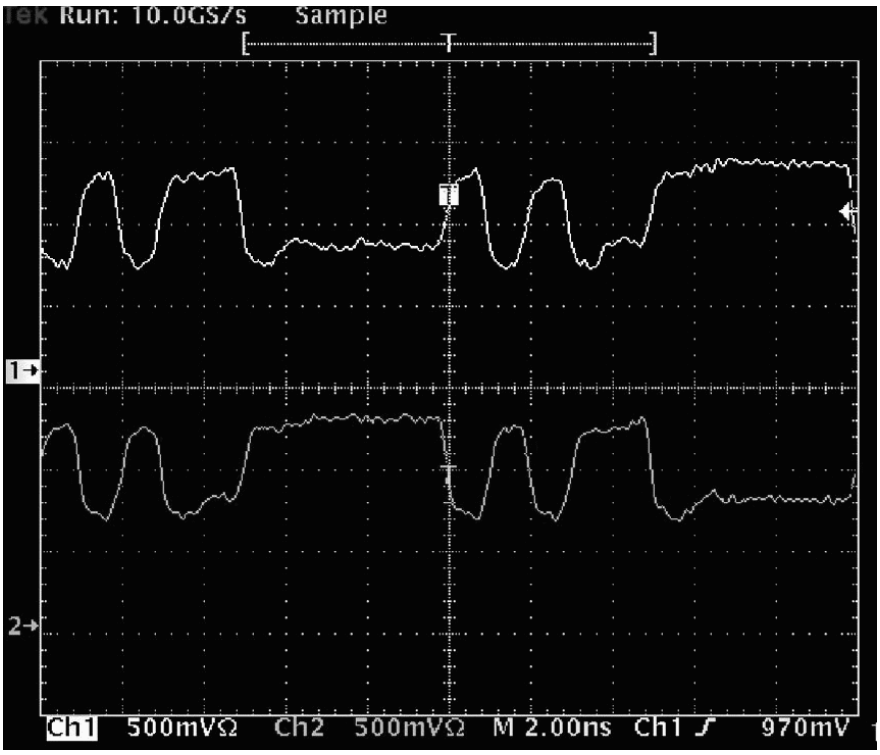


Figure 2-6. High-Speed data transmission at 1Gb/s

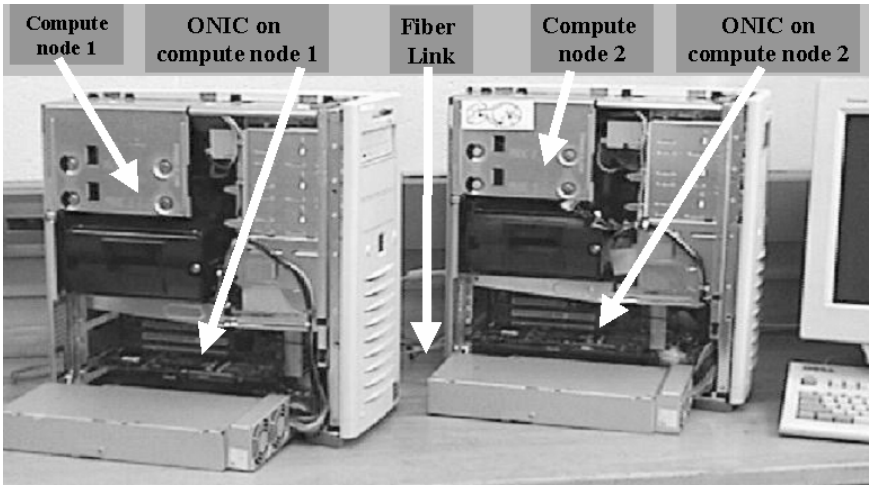


Figure 2-7. Compute servers communicating with ONIC through a 12-channel fiber ribbon

The application itself was simply the transmission of characters typed on the keyboard of one compute server to another. To start the transmission the compute server makes a request to the ONIC to establish communication. The ONIC in turn, sends K28.5 characters to establish the link and continues to send K28.5 characters to maintain the link. Each keyboard stroke is recognized by the driver, assigned a unique address (even if it is repeated) and sent to the ONIC through the PCI interface. The FPGA on the ONIC is programmed to recognize each character with its unique address. It sends it repeatedly to the SERDES until the next character with a different address is received for transmission. The SERDES 8b/10b encodes the data and sends it over the fiber link to the second compute server. Since the data bandwidth is very small compared to the bandwidth available, only one SERDES was used in this test. Figure 2-9 shows the steps involved in the message passing application test.

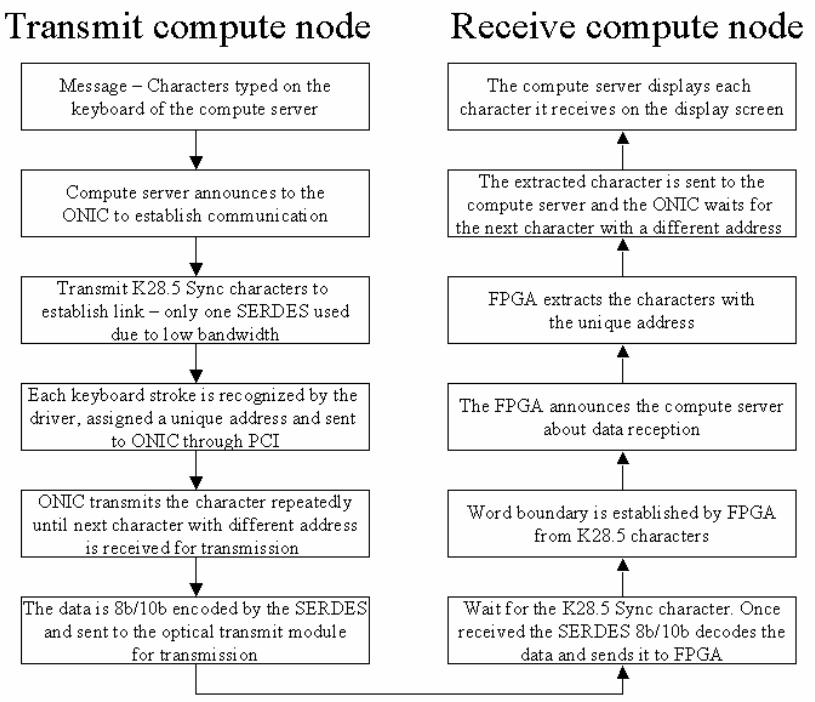


Figure 2-8. Message passing application test sequence

The SERDES on the ONIC of the receiving compute server decodes the data received. It recognizes the K28.5 characters sent to it and establishes a word boundary. The FPGA then waits for the data. As soon as it receives a new character with a unique address, it is sent to the compute server. The ONIC continues to monitor the received data. It sends the next character only when its address is different from the previous character received. Both the sent and received messages were displayed on the respective display screens of the compute servers. With only one SERDES used in the test the application was run at 4Gb/s bandwidth using only four channels in the 12-channel fiber ribbon.

3. MISALIGNMENT TOLERANCE IN THE RATS SYSTEM

3.A. Alignment Study Overview

For optical interconnections to be relevant to real systems they must be able to be manufactured and packaged inexpensively and robustly. This section discusses an optical design and packaging approach that utilizes multiple sizes (or scales) of optical elements to simplify the design of the optical interconnection and coupling while providing an enhanced degree of insensitivity to misalignments inherent in the packaging of these systems. The scales of the optical elements described are: the size of the integrated circuit (termed macro-optical), the size of the pitch of optical IO (termed micro-optical), and sizes in between (termed mini-optical) which are smaller than the size of the integrated circuit (IC) but cover several optical IO. This section describes the utility of elements of each of these scales and shows that through the combination of them simple robust systems can be constructed. Two case studies for applying this multi-scale optical design are examined. The first case study is a global chip-to-chip optical interconnection module. This multichip optical interconnection approach is termed *FAST-Net* after the contract that first developed the idea. In the RATS project, the *FAST-Net* approach is embodied to perform as the routing core of a multi-board optical interconnection fabric. This approach uses a macro lens array and mirror to implement the all-to-all optical interconnection pattern among an array of ICs on a single board. Micro- and mini-scale optical elements simplify the design of the macro-lens by performing corrections at scales where they are more effective. In this system over 11,000 optical links are implemented across a 5-inch multi-chip module with diffraction limited RMS (root mean square) spot sizes and registration errors less than 5 microns. This optical configuration (~11,000 links) was the initial design for the RATS module. The program proposed to leverage the concurrent effort of the VIVACE (VCSEL-based Interconnects in VLSI Architectures for Computational Enhancement) program to design a single optical system, which would be fabricated for both demonstrators (VIVACE and RATS). During the course of the RATS program, the availability of the Honeywell VCSELs and their integration became high-risk items. With the approval of DARPA, a backup technology choice was selected which would be much higher speed per link (~10x) with many fewer links (~1/10X). This revised optical demonstration module was based on Peregrine Semiconductor's Ultra Thin Silicon on Sapphire technology and was selected for the final prototyping. As there were not sufficient resources to perform two custom optical fabrications, and the revised prototype would be a subset of the possible number of optical sites running at much higher data rates, the optical system was kept as planned and under-populated by the Peregrine devices. The final optical module (capable of handling the ~11,000) optical I/O was experimentally evaluated across its entire field as described in Section 4. If the module were fully populated with optical links running at 3.125 Gbits/sec (as in the RATS demonstration) the bisection bandwidth of the resultant module would be ~17 Tbits/sec.

The second case study analyzes designs for board-to-board optical interconnections with throw-distances ranging from 1 millimeter to several centimeters. The RATS project built board-to-board interconnection prototypes with fiber based optical interconnections, but after design of the *FAST-Net* system was completed it was discovered that the misalignment tolerant aspects of the design would translate well into multi-board or connectorized approaches. In this case micro- and mini-scale optical interconnections provide insensitivity to misalignments. The results show the feasibility of an optical coupler that can tolerate the typical packaging misalignments of 5 to 10 mil without placing rigid constraints on the angular sensitivity of the modules. The multi-scale optical interconnection and coupling concept is shown to provide an approach to simplifying design and packaging – and therefore the costs – associated with implementing optical interconnection systems.

SPA chips with integrated VCSEL/detector arrays that have emitter and receiver elements sizes of 10 and 50 μm, respectively, and with element-to-element spacing as small as 125 μm, have been evaluated in a prototype interconnection fabric [29][30]. To fully exploit the smart pixel I/O density, the global optical interconnection module must provide flat, high resolution, near distortion-free image fields, across a wide range of ray angles, with low optical loss.

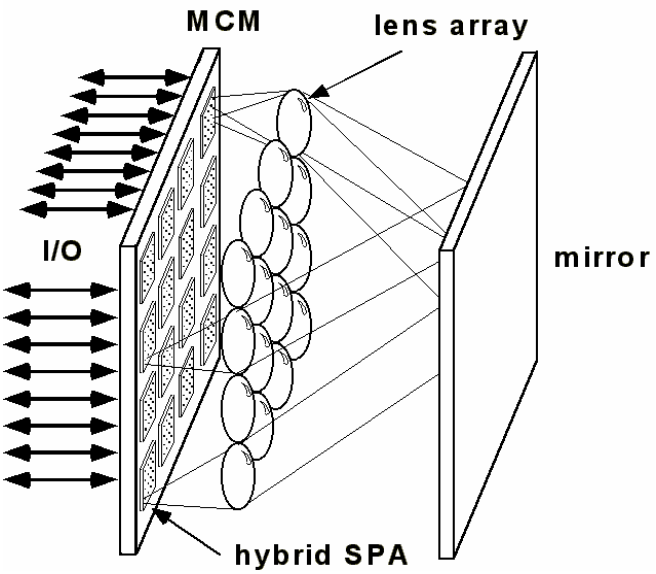


Figure 3-1. The multi-chip interconnection fabric achieves a high-density global multi-chip interconnection across an array of chips, thereby leveraging both the high bandwidth and high minimum bisection bandwidth ability of smart pixel technology and free-space optical interconnects

Modern optical design and manufacture techniques achieve wide-field imaging systems with high resolution. Low loss is achievable by optimizing lens designs that minimize the number of elements and employ antireflection coatings. Simultaneously achieving high registration accuracy across the entire field, however, is more challenging and can lead to complex multi-element solutions for the lens design. The initial *FAST-Net* prototype, which was developed before the RATS project, employed a set of matched 7-element off-the-shelf lenses. The prototype performed well in terms of registration and resolution [29][30] with SPAs that had small (<1 mm in diameter) VCSEL/photodetector clusters separated by several millimeters. However, this first generation prototype system was not suitable for RATS because it did not allow more closely spaced clusters, which are necessary for large-scale computer systems. We define registration accuracy here as the difference between the location of the image of a VCSEL and the location of its corresponding detector. In the system registration must be maintained at a level much less than the size of the detector (~50 μm) across the entire multi-chip plane (~10 cm). A comprehensive approach to designing the linking lens array, which maximally exploits the unique features of the *global multi-chip* VCSEL-based architecture, was required.

3.B. Case I: *FAST-Net* Global Multi-Chip Interconnection Module

In the design for the second generation *FAST-Net* prototype, there are 704 bi-directional channels on each of 16 SPAs, for a total of 11,264 FSOI links across the multi-chip module (MCM). There are therefore 16 spatially separated clusters of 44 VCSELs and 44 photodetectors on each SPA. The cluster is divided into spatially separate arrays of VCSELs and photodetectors. The shape of the cluster is octagonal, which results from the optimum circular shape as sampled by a regular square grid with a pitch of 175 μm. The circular apertures of the VCSELs used in the prototype are approximately 5 μm in diameter. The photodetectors have a dimension of 60 μm on a side. The maximum vertical/horizontal dimensions of the cluster are 1.575 mm. The 16 clusters are actually achieved by selectively utilizing VCSEL/photodetectors from a regular grid of VCSELs and detectors that are arrayed in a repeating pattern of 6 rows of VCSELs, 6 rows of photodetectors, and 1 row of unused elements all on a 175 μm grid. The clusters are formed by sampling adjacent sets of 5 rows of VCSELS and photodetectors to achieve the desired cluster configuration. The optical arrays are area bump bonded to a matching array of driver and receiver circuits on the underlying silicon SPA IC. The distance between the centers of adjacent clusters in on the SPA is 2.275 mm and therefore the overall SPA chip size is ~10 mm.

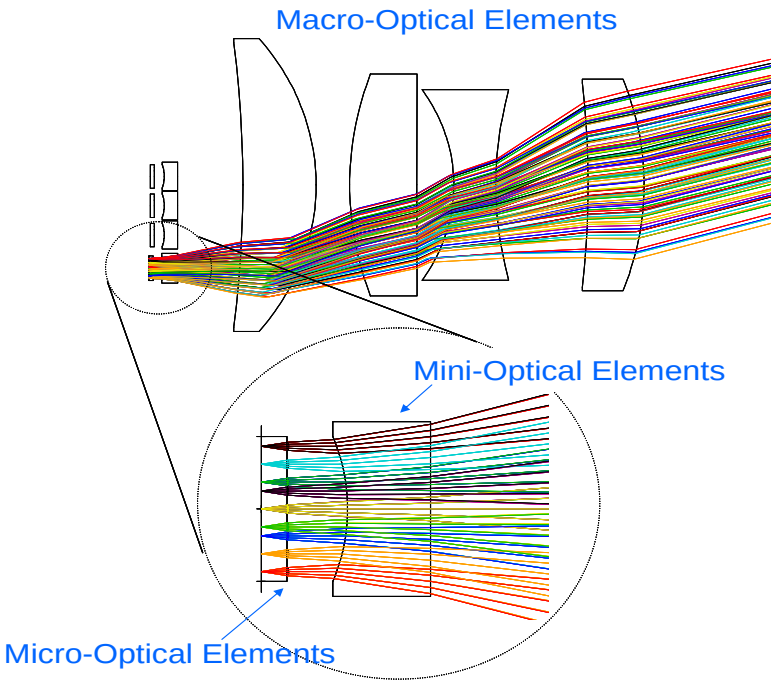


Figure 3-2. Multi-scale optical interconnection design for the global multi-chip system

The registration and resolution design goals for the second generation *FAST-Net* optical system prototype derive from an overall goal of 90% light capture efficiency at the detector, meaning that the blur spot of any VCSEL image should be confined and registered so that its corresponding 60 μm wide square detector captures 90% of the light energy. This level of performance will ensure that the receiver detects sufficient light from the VCSEL and optical crosstalk between adjacent channels will be negligible. The combined levels of distortion error and blur size should be small enough to ensure this level of performance. Minimizing the number of elements in the overall lens and employing antireflection coatings to minimize reflection losses should maximize the overall optical transmission efficiency for the optical system. To minimize the overall size of the MCM and achieve good SPA chip real estate utilization, a maximum lens diameter of 3 cm and an f-number of less than 1.25 were desired.

The overall goal of the design is to implement the required global interconnection pattern across a 4x4 chip array while minimizing the complexity (i.e., number of elements, cost, etc.) of the optics. Figure III-2 is a schematic depiction of one of the 16 custom-designed lenses in the system. It consists of three distinct types of optical elements referred to as "micro" (1 per VCSEL or detector), "mini" (1 per cluster of parallel VCSEL links), and "macro" (1 per SPA). In this design the micro-optical elements are solely responsible for reducing the numerical aperture (divergence angle) of the VCSEL beam, thereby reducing the overall required lens complexity as discussed below. The mini-optical elements effect a distortion eliminating beam-steering function that was recently proposed [31] and evaluated [32][33][34]. The concept uses fixed mini-optical beam-steering elements to achieve symmetrical, and hence distortion-free, ray paths through the global optical interconnection system. This approach exploits the inherent small numerical aperture (NA) of VCSELs to eliminate distortion by achieving holo-symmetry for each pair of lenses. The macro-optical elements (4 in each lens) implement the global optical interconnection pattern. The complexity of these macro-elements is greatly reduced by the presence of the micro-optical and mini-optical elements. The macro- and mini-elements are contained in a single barrel as shown in Figure 3-3. The micro-elements, which reduce the numerical aperture of the VCSEL beam and therefore simplify the remainder of the lens design [35][36], are integrated directly on the VCSEL/detector array, via mounting to a transparent superstrate, as depicted in the blow-up in Figure 3-2.

Figure 3-3. Lens barrel containing multi-scale optical elements for global multi-chip interconnections. Macro-optical elements are the size of the barrel, whereas an array of mini-optical elements is mounted to an optical flat in the barrel

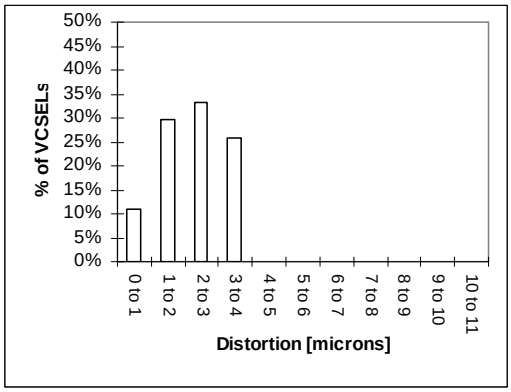


Figure 3-4. Nominal mis-registration (distortion) for multi scale global multi chip lens design

3.B.1. Lens Performance

The multi-scale lens design effectively partitions the critical VCSEL cluster imaging requirements into: numerical aperture control, beam steering, and off-axis imaging. The combination of the micro-, mini-, and macro-scale optical elements provides an effective solution for the stringent optical system requirements. The spot sizes for all of the links are 3-4 microns. Figure 3-4 shows a histogram of distortion for each VCSEL/detector link. The multi-scale lens design corrects distortion from 8% (optimized without mini-lens elements) to less than 0.08%. This greater than 100x reduction in distortion reduces misalignments from 580 microns to <4 microns. This lack of distortion is highly unusual for off-axis imaging systems and it is achieved through the beam steering (mini-scale optical elements) of the low resultant NA VCSELs (created by the micro-scale elements). Without the combination of these three scales of optical components a lens would be unreasonably complex in order to meet the wide field global off-axis imaging requirements dictated by the system. The design presented in this section enables global optical interconnection modules, such as the one depicted in Figure III-1, to fully exploit the anticipated Terabit/sec/cm^2 capabilities of smart pixel technology.

3.B.2. Misalignment Tolerance

Since the multi-scale optical interconnection system (half of which–for any pair of chips–is depicted in Figure III-2) is implemented as two infinite-conjugate-ratio systems in an imaging configuration, one would expect misalignments of the lens barrels to directly translate into misalignments of image spots. However, this is not the case as the multi-scale design provides a measure of immunity to lens misalignments. Recall that in the design described above, the mini- and macro-optical elements are mounted in the same barrel, whereas the micro-optical elements are directly integrated onto the superstrate of the optoelectronic devices. As the lens barrel is translated, due to some source of alignment or operational environment error, only the mini- and macro-optical elements are displaced. Table 3-1 compares the performance of the nominal system (no misalignments) to systems with displacements and rotations. Column 3 represents the data from a 10 micron displacement of the lens barrel. Note that the image location (measured by distortion) remains relatively unchanged, where it would have been expected to translate a corresponding 10 microns. Also, note that the spot sizes have only increased slightly–up to a ~8-micron radius. Since the image is still well centered on the detector, the larger detector area will readily capture the energy. Column 4 represents the equivalent data for a 20-micron displacement of the lens barrel. In this configuration, the distortion is still negligible, but the spot size for some of the links is beginning to be problematic. These links are at the edge of the cluster and under this displacement, are hitting near the edge of optical elements. Systems requiring additional displacement accuracy could be designed with smaller clusters allowing more margins at the lens edges. Placing the mini-optical elements in the lens barrel breaks the rotational symmetry of the macro-optical elements so rotational misalignments of the barrels must be considered as well. Column 5 of the table represents similar results for a 1-degree rotational misalignment of the lens barrel. The symmetry provided by the beam steering of the mini-optical elements provides a well-balanced point about which the effects of misalignments are mitigated by the use of micro-optical scale elements.

Table 3-1. Misalignment performance of multi-chip interconnection module (measurements in microns)

	Nominal	10 um	20 um	1 deg
Min RMS Spot size	3.36	3.36	4.14	3.67
Max RMS Spot size	3.64	8.02	23.84	5.75
Mean RMS Spot size	3.49	4.56	11.56	4.18
Min Distortion	0.7	0.53	0.47	0.96
Max Distortion	3.9	4.68	5.01	6.32
Mean Distortion	2.33	2.69	2.52	2.63

3.C. Case II: Board-To-Board Optical Interconnection

Although the multi-scale optical approach was originally developed for global chips-to-chips optical interconnection modules, its misalignment insensitivity in that domain makes it an interesting candidate for the RATS board-to-board optical interconnection approach. The RATS board-to-board demonstration was based on parallel fiber optic cabling, but the misalignment tolerant aspects of the free-space approach described above directly impacts the board-to-board implementation. Even if the final design for board-to-board interconnections dictates the use of parallel fiber optical cabling, the following approach may greatly simplify the connectors by lessening the manufacturing and packaging tolerances of the board to fiber interface. An interconnection application in which an array of emitters is linked to an array of detectors or guided-wave channels over a short (1 mm–2 cm) throw distance is both of great interest and is plagued by the effects of misalignments which results in an increased packaging cost. In such a configuration, the lack of "global" interconnections causes degeneracy between the scales of mini-optical and macro-optical elements, i.e., often the cluster size is the same as the array size. When this is the case, we use the more commonly used macro-optical designation to describe the scale of the element. In order to quantify the benefits of the multi-scale design approach we compare it to macro-optical-only approach. Micro-optic approaches have been studied in detail but are limited in throw distance and do not provide the tradeoff between angular and translational misalignments which we will show in the multi-scale approach. In this analysis, both approaches image an array of VCSELs (with a 3 mm field) onto an associated array of detectors. The analysis assumes that the link is broken into two halves: transmitting plane with its associated optics and receiving plane with its optics. All misalignments happen between these two halves.

3.C.1. Macro-Optical Approach

The first optical interconnection approach evaluated was a macro-optical one. In this case, the size of the optical elements will be on the order of the optical array size (e.g., several millimeters). An expanded beam (infinite conjugate ratio) interconnection between planes will be utilized. This optical interconnection approach provides maximum tolerance to misalignments in the x and y directions (i.e., perpendicular to the optical axis (z)), as shifts between the two system halves do not affect the angle of the beam between them. The increased tolerance to x and y misalignments comes at a price of increased sensitivity to angular misalignments. We can bound the best performance of such a system by considering the lenses to be perfect elements (i.e, there is a tangential relationship between angle and position). While the x and y tolerances are on the order of the lens radius, the tolerances to angular rotations out of the plane (θ) and within the plane (φ) are extremely tight:

$$\delta x_{\theta} = f\,(\tan(\operatorname{atan}(x/f) + \delta\theta_x)) - x, \qquad [3\text{-}1]$$

and

$$\delta x_{\varphi} = x * \varphi \qquad [3\text{-}2]$$

where δx is the displacement of the image due to the misalignment (in either θ or φ, f is the focal length of the macro-lens, x is the radial off-axis distance of the VCSEL, $\delta\theta_x$ is the angular rotation out of the plane in the direction of the x axis, and φ is the rotation within the plane (about the optical axis). Figure 3-5 depicts a typical macro-optical interconnection approach when perfectly aligned (top) and under a 250 micron y-displacement, 250 micron z-displacement and 1 degree angular misalignment (bottom).

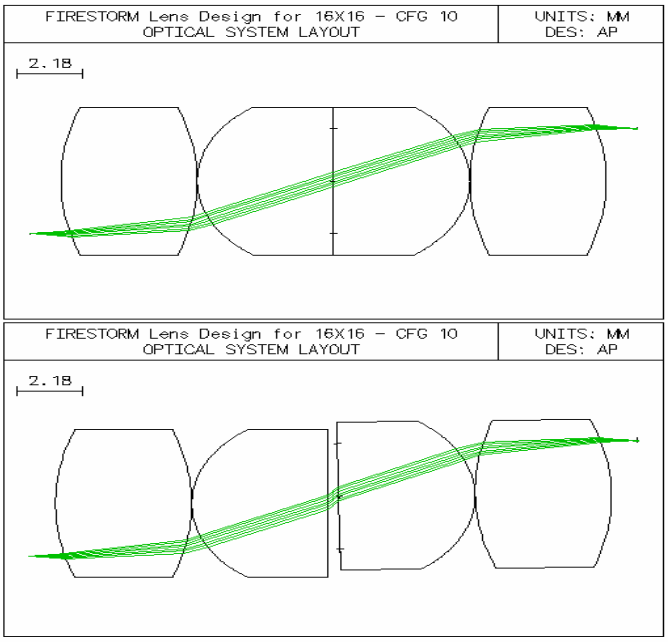


Figure 3-5. Macro-optical interconnection perfectly aligned (top) and with 250 micron displacement in the plane, 250 micron displacement along the optical axis and 1 degree rotational (out of plane) displacement (bottom)

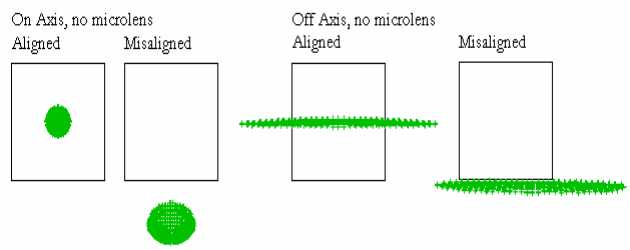


Figure 3-6. Spot diagram of on-axis point and off-axis point, aligned and misaligned for macro-optical interconnection. Square depicts boundary of the 75 micron side photodetector

Figure 3-6 shows associated predicted spot diagram for on-axis and full field object points in the aligned and misaligned systems in relation to a 75 micron detector width. As the figure shows, slight rotations between the planes would cause link failure in a macro-optically interconnected system.

3.C.2. Multi-Scale Micro-Macro-Optical Interconnection Approach

In the multi-scale approach macro-optical interconnection lenses and micro-optical elements are combined with the hope of achieving the similarly increased system performance to the global multi-chip system. By including micro-optical elements in a macro-optical expanded beam interconnection approach the overall sensitivity to angular misalignments can be reduced. This comes at the expense of increasing the positional (x, y) insensitivity inherent in the macro-optical approach. The goal, therefore, is to design the interconnections optics to provide the best trade-off in angular and positional tolerances as determined by the application packaging requirements and constraints. As before, in assuming a perfect lens element, a bound on the sensitivities to positional and angular misalignments of the multi-scale approach yields:

$$\delta xr=(d1-d2)*(\delta r/f), \qquad \text{[III-3]}$$

$$\delta x\theta=(d2/d1)*[f\,(\tan(\operatorname{atan}(x/f)+\delta\theta x)-x], \qquad \text{[III-4]}$$

$$\delta x\varphi=(d1-d2)*[x*\varphi]/f, \qquad \text{[III-5]}$$

where d_1 is the distance from the micro-lens to the old image plane, d_2 is the distance from the micro-lens to the new image plane, δr is the lateral shift of the lens system (misalignment), and θx_r is the shift in image position due to δr. Figure 3-7 depicts the resulting mis-registrations due to the various misalignments. The top of the figure represents those of the macro-optical-only system, whereas the bottom of the figure represents those of the multi-scale micro/macro-optical approach.

Note that the macro-optical approach does not suffer under small translational misalignments (upper left) but is sensitive to angular misalignments. Some of this translational insensitivity is traded off for angular sensitivity in the multi-scale approach. The terms in the brackets of equations 4 and 5 are the previous results for the macro-optical interconnection approach. Notice that the shift of the image plane due to the presence of micro-optical elements yields a direct (and inverse) tradeoff between sensitivities to misalignments due to angles within and out of the image plane. Figure 3-8 depicts a typical

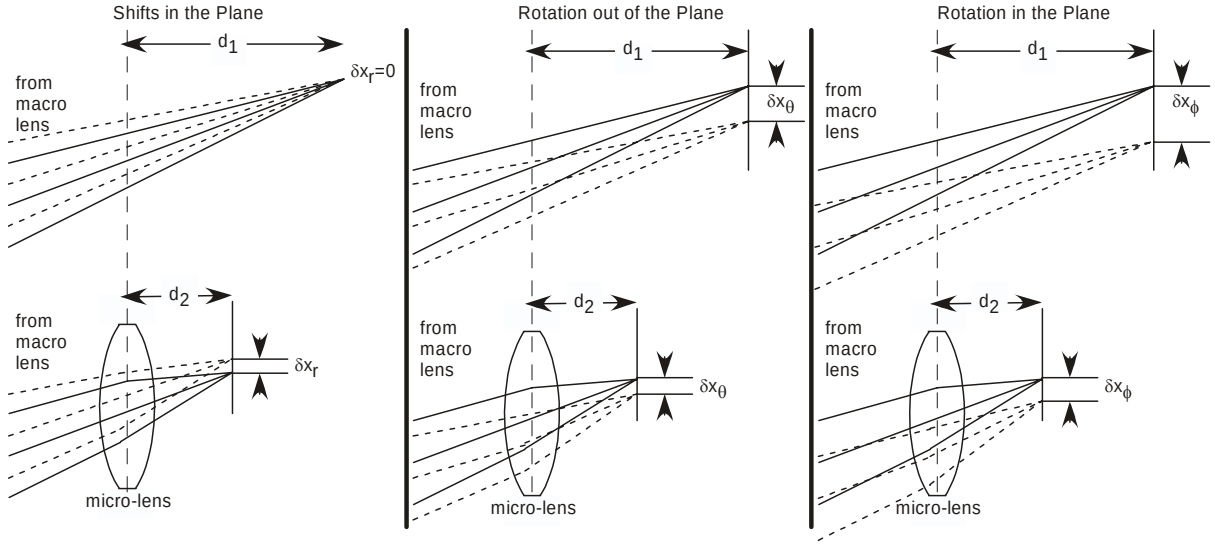


Figure 3-7. Schematic diagram depicting mis-registrations due to misalignments in macro-optical only (top) and multi-scale (bottom) approaches

multi-scale macro-optical interconnection approach when perfectly aligned (top) and under a 250 micron y-displacement, 250 micron z-displacement and 1 degree angular misalignment (bottom). Figure 3-9 shows associated spot diagram for on-axis and full field spots in the aligned and misaligned systems in relation to a 75 micron detector. As the figure shows, the slight rotations, which plagued the macro-optical approach are readily handled by this system.

The multi-scale micro-macro-optical interconnection approach allows for a fluid trade space between sensitivities in positions and rotations between planes. It main limitation is in its macro-scale: if the throw distance is reduced to an extremely small distance then the focal lengths of the macro-optical elements will necessarily become very short. Combining this with a manufacturing constraint of avoiding lenses with an excessively low f# would limit the field size of the interconnection (and therefore the number of links behind the macro-lens). For throw distances of 1 cm or more the multi-scale approach will work well. For smaller center-to-center spacings the micro-optical interconnection approach may be more practical and would be allowable as the diffraction limits would not hinder them in this domain.

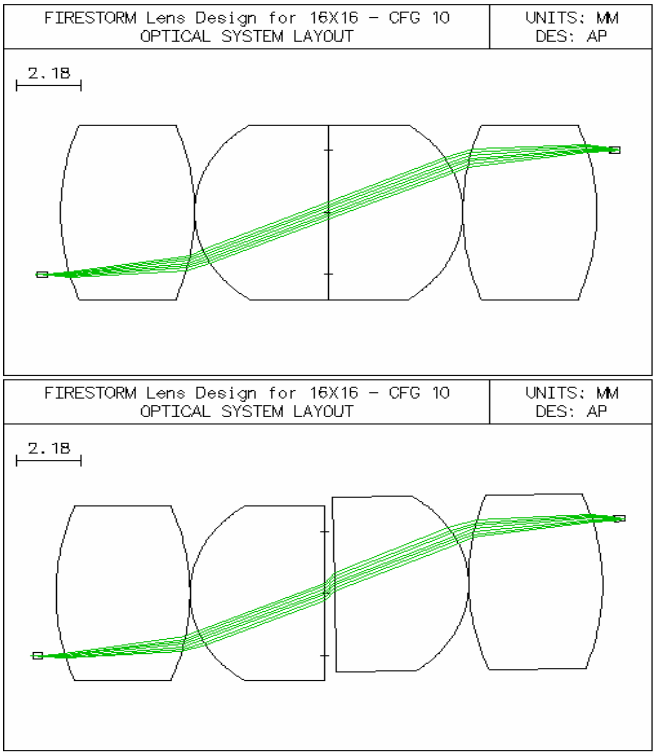


Figure 3-8. Multi-scale optical interconnection perfectly aligned (top) and with 250 micron displacement in the plane, 250 micron displacement along the optical axis and 1 degree rotational (out of plane) displacement (bottom)

3.D. Free-Space Optics Summary

This section introduced a hybrid optical design and packaging approach that utilizes multiple sizes (or scales) of optical elements to simplify the design of the optical interconnection and coupling while providing an enhanced degree of insensitivity to misalignments inherent in the packaging of these systems. The utility of elements of each of these scales was described, and it was shown that, through the combination of them, simple robust systems can be constructed. This section examined two case studies directly related to the RATS/PCA program domain in which multi-scale optical design was applied. The first case study involved a global chips-to-chips optical interconnection module for the RATS intraboard switch/router demonstration, which uses a macro lens array and mirror to effect the all-to-all optical interconnection pattern among an array of ICs on a single board. Micro- and mini-scale optical elements were shown to simplify the design of the macro-lens by performing corrections at scales where they are more effective. In this system over 11,000 optical links are implemented across a 5 inch multi-chip module with diffraction limited RMS spot sizes and registration errors less than 5 microns. The second case study analyzed designs for PCA interboard optical interconnections with throw-distances ranging from 1 millimeter to several centimeters. In this case micro- and macro-scale optical interconnections provide insensitivity to misalignments. The results show the feasibility of an optical coupler that can tolerate the typical packaging misalignments of up to 250 microns without placing rigid constraints on the angular sensitivity of the modules. Multi-scale optical interconnection and coupling design were shown to provide an approach to simplifying design and packaging, and therefore the costs, associated with implementing optical interconnection systems.

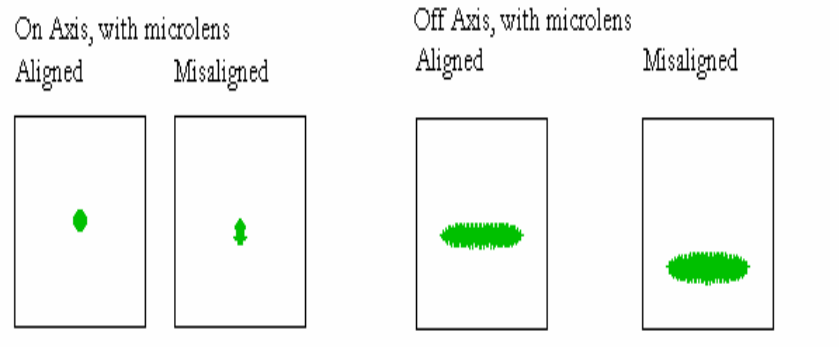


Figure 3-9. Spot diagram of on-axis point and off-axis point, interconnection. A comparison with Figure III-8 shows the benefits of the hybrid approach in reducing mis-registrations aligned and misaligned for multi-scale optical

4. EXPERIMENTAL FREE-SPACE MODULE EVALUATION

4.A. Demonstration Overview

Figure 4-1 depicts the FPGA and smart pixel chip system that are fully interconnected by the global optical fabric. Four hybrid optical smart pixel chips are attached to the board via fuzz-button cinch connectors. The smart pixel chips combine Ultra-Thin Silicon (UTSi) ASICs with VCSEL and detector arrays and micro-lens collimating elements. The inherent transparency of the sapphire substrate of the ASIC obviates the need for any substrate removal of the VCSEL and detector arrays [37]. The emitted light passes through the sapphire substrate and the back surface of the sapphire provides a convenient location to implement beam-conditioning micro-optics. A multi-scale multi-element lens will be placed above each of the smart pixel chips. The resultant system is folded by placing a mirror above the entire assembly. The UTSi ASIC, shown in Figure 4-2, is based on Peregrine Semiconductor's FOCUTSpak 1-D parallel fiber transceivers. The FOCUTSpak is a 4 x 3.125 Gbps transceiver implemented with alignment pins for an MT connected fiber ribbon cable (MT is a common connector standard that originally stood for "mechanically transferable"). Since the demonstration module discussed in this section interconnects 4 smart pixel chips each with 4 channels, the circuitry of the single 1x4 transceiver was replicated 4 times within a single smart pixel assembly. In this way each smart pixel chip contains 16 VCSELs and 16 detectors (4 groups of 4 elements) each running at multi-gigabit data rates. The goal of the demonstration is to have each smart pixel assembly interconnected pairwise with a bi-directional 10 Gbps "pipe". The resultant aggregate data is 160 Gbps (4 smart pixel assemblies x 4 groups of I/O x 4 links x 2.5 Gbps). In this system, a pair of Virtex II Pro FPGAs will be interconnected to each smart pixel assembly through a printed circuit board. The Virtex II Pro family of FPGAs was chosen because of the multi-gigabit transceivers, serialization, and deserialization, present in this family. The parts that were used have a total of 8 multi-gigabit transceiver ports per FPGA; so 2 FPGAs are required to attach to each smart pixel assembly. An additional benefit of the Virtex II Pro FPGAs is the presence of a PowerPC core which will allow for some data processing interconnected by the high speed free-space fabric.

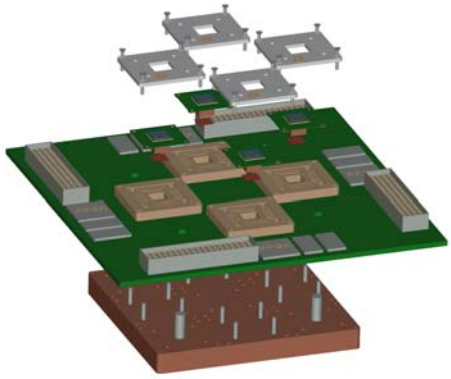


Figure 4-1. Board and baseplate system for aligning hybrid smart pixel arrays in system

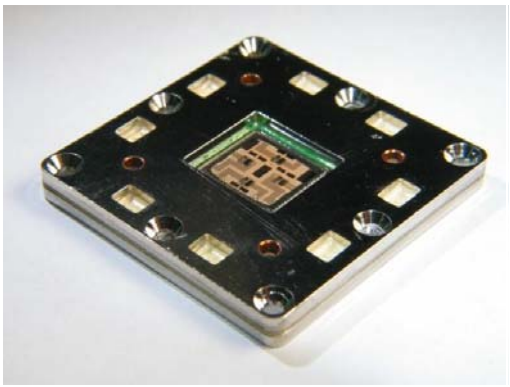


Figure 4-2. Peregrine semiconductor's modified FOCUTSpak with 4 1x4 VCSEL arrays and 4 1x4 photodiode arrays

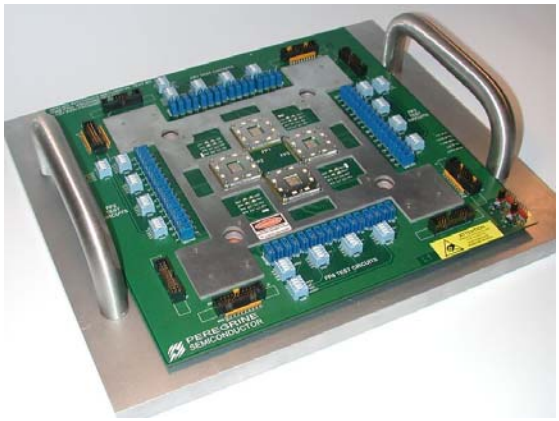


Figure 4-3. Photograph of assembled alignment baseplate and FOCUTSpaks. The optical module has been removed

Figure 4-3 is a photograph of the assembled electronics and baseplate with the optical assembly removed. Each FOCUTSpak is screwed down onto the baseplate after having been aligned to registration pins. All alignment sensitive features are referenced to this same baseplate so that there is no stackup of registration error. This approach assured accurate positioning of the active optical devices with respect to one another, leaving most of the error budget to the alignment of the lenses themselves. Since the multi-scale optical interconnection system (half of which–for any pair of chips–is depicted in Figure III-2) is implemented as two infinite-conjugate-ratio systems in an imaging configuration, one would expect misalignments of the lens barrels to directly translate into misalignments of

image spots. However, this is not the case as the multi-scale design provides a measure of immunity to lens misalignments. Recall that in the design described above, the mini- and macro-optical elements are mounted in the same barrel, whereas the micro-optical elements are directly integrated onto the superstrate of the optoelectronic devices. As the lens barrel is translated, due to some source of alignment or operational environment error, only the mini- and macro-optical elements are displaced.

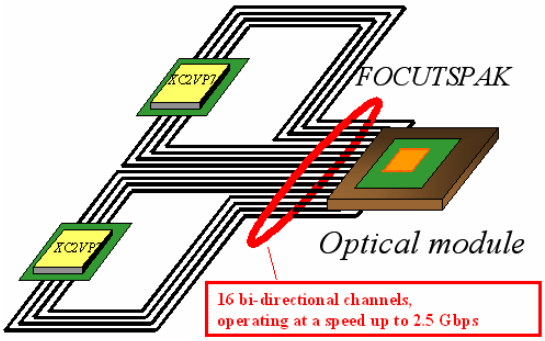


Figure 4-4. One port of the system with 16 channels using 2 FPGAs–there are four such ports on the board

4.B. Test Bed Specifications

The RATS system is made up of four nodes, with each node containing a FOCUTSpak optical module and two Xilinx Virtex II Pro FPGAs, as seen in Figure 4-4. Each FOCUTSpak has 16 outputs and inputs, making for a total of 64 bi-directional channels. The ports are connected via free-space optical channels. With each channel running at the maximum 2.5 Gbps, the aggregated external I/O throughput of the system is 160 Gbps. Connected to each FOCUTSpak are two FPGAs. These are used to generate data to be sent via the optical module as well as receive data from the modules, as seen in the data flow example in Figure 4-5 (where the yellow lines represent electrical data and the orange lines represent optical data). Additionally, the FPGAs are used to program the registers that set the bias and modulation currents for the VCELS. The Xilinx Virtex II Pro FPGA was chosen because of its integrated multi-gigabit transceivers (MGT), Figure 4-6. Each FPGA has eight MGTs, each capable of running up to 2.5 Gbps with the chosen package type. With all eight FPGAs running in the system one can sustain the peak system throughput of 160 Gbps.

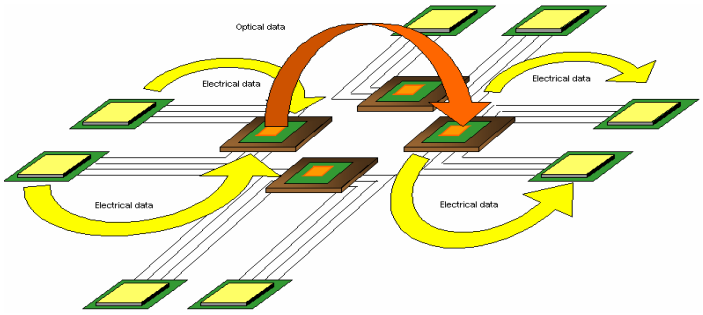


Figure 4-5. Data flow diagram of the four-port system

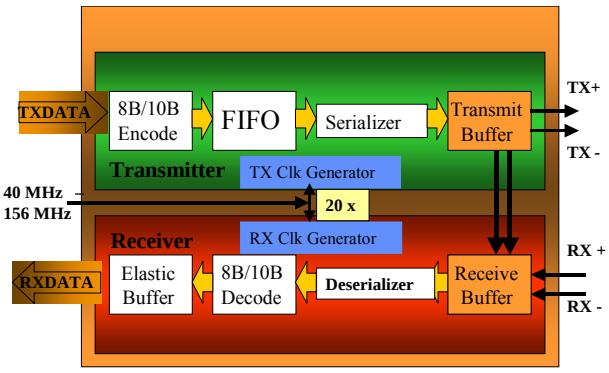


Figure 4-6. Block diagram of the Xilinx Virtex II Pro Rocket I/O Technology (MGT). Operating speed: 622Mbps to 3.125 Gpbs

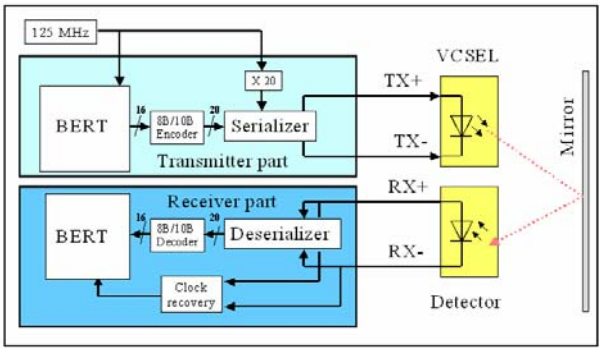


Figure 4-7. Block diagram schematic of the test bed for the RATS system

To characterize the optical link integrities, a bit-error rate tester (BERT) was designed and implemented on the FPGAs. A BERT is composed of a transmitter and a receiver (Figure 4-7). The transmitter side first sends a synchronization vector and then will continuously send a pseudorandom bit-stream. On the receiver part, once the synchronization vector is seen, the same pseudorandom bit-stream is internally generated. The received bit-stream is then XORed with the internally generated stream. Any ones that occur from an XOR are registered as an error and a counter is incremented. If run for a long enough period of time, one can characterize the bit error rate for the link under test. The BERT was tested on the system using electrical only links to verify that the test runs correctly.

4.C. Alignment and Testing of the Optical System

The electrical characteristics of the RATS module were verified and the optical alignment in the RATS module was conducted using the lenses designed for the VIVACE system. The electrical operation of the board was tested to confirm that the correct signals appeared on the correct output pads. This was to ensure that the FOCUTSpaks would be receiving the correct and required driving signal inputs. Once the electrical characteristics of the RATS PCB were tested thoroughly, the board was sent to Peregrine Semiconductor, where the four FOCUTSpaks were integrated with the board, allowing the optical alignment procedures to begin. After taking an inventory of the working optical I/Os, the optical alignment process was begun, with limited success. Direct observation showed that one lens was aligned to itself, so that the VCSELs from the on-axis cluster pass through the optical system and appear to impinge upon the corresponding detectors (figure below), but no electrical signal was detected at the output of the receiver circuit. It was not entirely clear if the detectors were receiving enough light, or if there is an issue with the receiver circuitry preventing a clear output signal. This question was explored by troubleshooting the system, with the aim of demonstrating multiple links between multiple chips.

Several steps were taken to verify the optical alignment of the VIVACE lenses in the RATS system. One key issue that inhibited the ease of demonstration was that the FOCUTSpaks appeared to experience connector issues, as the VCSELs and detectors showed intermittent and unpredictable operation. There also appeared to be an issue with the receiver circuitry, but the optical alignment was verified after adjusting a number of the electrical signal levels. For example, when each of the VCSELs in a cluster was illuminated in turn, it would cause a signal to appear on the corresponding detector pad when probed, but if multiple VCSELs in an array were turned on simultaneously, at best one detector in the destination cluster would show reception of the signal. This pointed to electrical cross talk or an issue with the receiver circuitry, but it did not rule out the possibility of optical cross talk. The question of optical cross talk was ruled out by two methods. First, from direct observation it was clear that the light impinging upon each detector was well collected in a small spot size. Second, it was shown that the light between the focused beam spots did not contribute to a detectable level. The focused beams were observed using a fiber-coupled detector which, when moved between the VCSEL spots, detected the transmitted signal only when the aperture was very closely aligned with the beam spot, as shown in Figure 4-9.

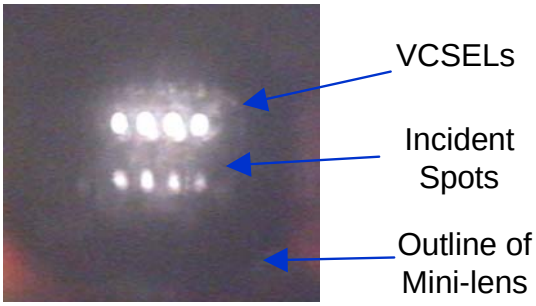


Figure 4-8. Direct observation of the VCSELs and light impinging on detectors for the on-axis cluster

Once this was observed, the optical links were tested one by one, so that a total of 22 of the 64 possible links were verified. Among these 22 links, all three possible combinations of lens pairs were represented (self, opposite, adjacent). Once the simultaneous optical alignment between 22 VCSEL-detector pairs was demonstrated, the misalignment tolerance was tested by deliberately perturbing the system with misalignments in several orientations. While the mini-lenses disrupt the axial symmetry of the macro-lenses, slight axial rotation did not contribute to noticeable misalignment. Likewise, lateral perturbation in the plane of the optical I/O was tolerable up to about the diameter of the microlenses, as predicted by simulation. The most important alignment consideration appeared to be the axial tilt of the macro-lens with respect to the plane of the optical I/O. This critical kind of alignment was ensured by the fact that the lens barrels were evenly seated on a custom-machined base-plate, which was leveled with respect to the optical I/O. A demonstration of the optical alignment and the misalignment tolerance of the system were performed for a panel of observers from DARPA, wherein, probing the detector pads and viewing the

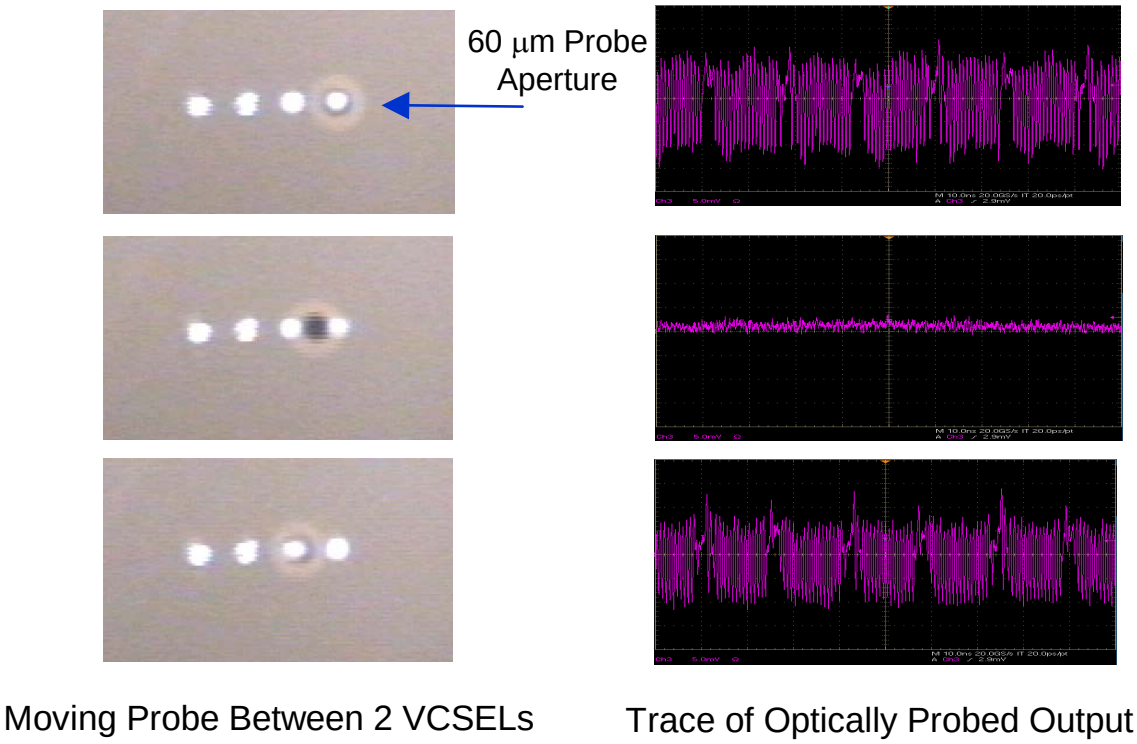


Figure 4-9. Demonstration of the absence of optical cross talk in the system

output on a precision oscilloscope showed several of the links. The misalignment tolerance was likewise demonstrated by showing that the probed signal remained on the oscilloscope while the lenses were translated over a significant distance. Once the demonstration was completed, the system was disassembled and the board was shipped to Peregrine Semiconductor so that the connector issues of the FOCUTSpaks could be addressed.

4.D. Experiment Summary

High-speed free-space optical interconnects were demonstrated to interconnect four OE-enabled integrated circuits. Data transmission rates of 2.5 Gbps were observed, but no working links were demonstrated, due to connector issues with the OE chips. However, 22 of 64 possible optical interconnect paths were shown to be simultaneously achieved with relative ease and a useful amount of misalignment tolerance.

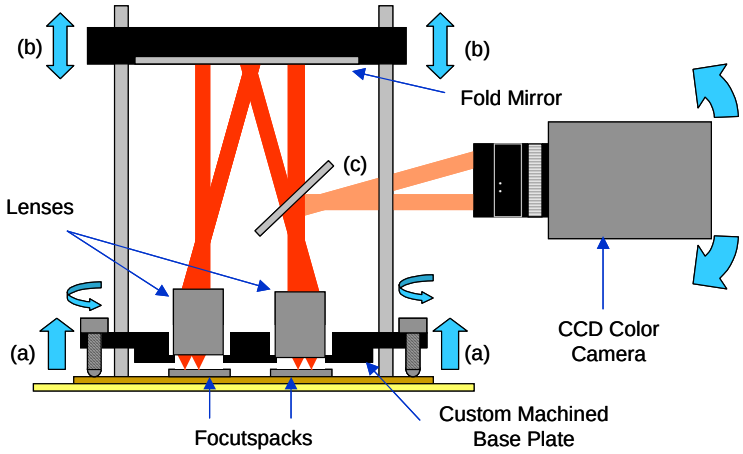


Figure 4-10. Schematic showing the alignment procedure

5. SOURCE-SYNCHRONOUS DOUBLE DATA RATE (DDR) PARALLEL OPTICAL INTERCONNECTS

5.A. DDR Parallel Optical Interconnect Overview

Double data rate (DDR) signaling is widely used in electrical interconnects to eliminate clock recovery and to double communication bandwidth. This section describes the design of a parallel optical transceiver integrated circuit (IC) that uses source-synchronous, DDR optical signaling. This parallel optical transceiver IC was developed to demonstrate fast power-efficient links that could be used as integrated communication channels integrated with high-speed microprocessors. The integration of communication links with a high-performance microprocessor is outside the scope of the RATS project, but is critical for the adoption of photonics technology in PCA systems. On the transmit side, two 8-bit electrical inputs are multiplexed, encoded and sent over two high-speed optical links. On the receive side, the procedure is reversed to produce two 8-bit electrical outputs. Our IC integrates analog VCSELs, drivers and optical receivers with digital DDR multiplexing, serialization, and deserializaton circuits. It was fabricated in a 0.5-micron Silicon-on-Sapphire (SOS) CMOS process. Linear arrays of quad VCSELs and photodetectors were attached to our transceiver IC using flip-chip bonding. A free-space optical link system was constructed to demonstrate correct IC functionality. The test results show successful transceiver operation at a data rate of 500 Mbps with a 250 MHz DDR clock, achieving 1 Gbps of aggregate bandwidth. While our DDR scheme is well suited for low-skew fiber-ribbon, free-space, and waveguide optical links, it can also be extended to links with higher skew with the addition of skew-compensation circuitry. To our knowledge, this is the first demonstration of parallel optical transceivers that use source-synchronous DDR signaling.

Typical I/O architectures transmit a single data word on each positive or negative clock edge and are limited by the associated clock speed. For example, to achieve a 400-megabit per second (Mbps) transfer rate, a system requires a 400-MHz clock. Many new applications have introduced DDR I/O signaling to improve upon single data rate (SDR) signaling, because it allows for higher throughput. While SDR captures data on one edge of a clock, DDR captures data on both edges of the clock, doubling the throughput for a given clock and accelerating performance. Thus a 200-MHz clock can capture a 400-Mbps data stream. In this manner, we can achieve the same performance with lower power consumption.

Many emerging I/O standards, such as HyperTransport [38], RapidIO [39] and POS-PHY Level 4 [40] employ source-synchronous DDR schemes in their link protocols. Figure V-1 shows the physical link interface between two HyperTransport devices [38]. Two point-to-point unidirectional links consisting of CLK (clock), CAD (command, address and data) and CTL (control) are used between two HyperTransport devices. The clock signal is sent along with data to reduce global clock skew. Low voltage differential signaling and DDR clocking are used on the CLK and CAD lines. Incorporating DDR signaling into our optical transceiver design, we have implemented a two-channel parallel optical transceiver IC using a source-synchronous DDR scheme similar to that used in the HyperTransport links. Analog VCSEL drivers and optical receivers are integrated with digital DDR multiplexing, serialization and deserializaton circuits in our transceiver IC. On the transmit side, two 8-bit electrical inputs are multiplexed into two DDR serial streams and sent over two high-speed optical links through VCSEL driver circuits. The clock signal, oscillating at half the serial data rate, is transmitted along with each data channel. On the receive side, the optical serial data are received by photodetector receiver and passed to DDR flip-flops that sample the data on the rising and falling edges of the clock signal. The serial data are then further demultiplexed into parallel data.

The IC was fabricated with 1x4 VCSELs and 1x4 photodetectors heterogeneously bonded to the CMOS circuitry. To test the functionality of the IC, we built a free-space optical-link demonstration system between two chip-carrier boards separated by 76.2 mm on a PCB main board. Both electrical and optical tests have been performed on the IC. The test results show that the transceiver IC and the optical links are fully operational at a data rate up to 500 Mbps with 250MHz DDR clock per channel, achieving a gigabit of aggregate bandwidth. The optical links we used for demonstration suffered minimal skew, but the DDR scheme can be extended to links with higher skew with the addition of skew-compensation circuitry [41].

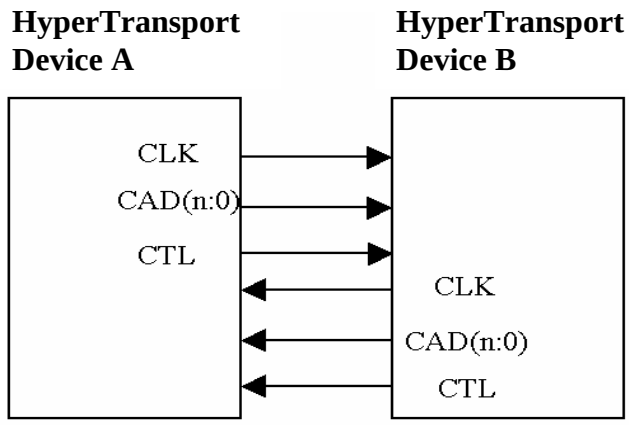


Figure 5-1. HyperTransport IO link. Commands, addresses, and data (CAD) all share the same bits. CADs can be 2, 4, 8, 16, or 32 bits wide. Each data path includes a control (CTL) signal and one or more clock (CLK) signals. The CTL signal differentiates commands and addresses from data packets. For every grouping of eight bits or less within the data path, there is a forwarded CLK signal

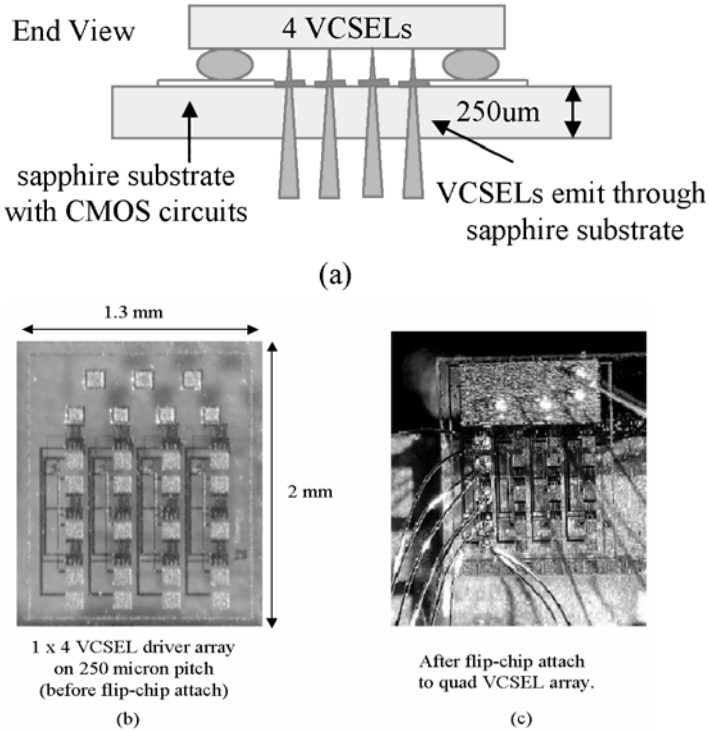


Figure 5-2. (a) End view of VCSEL flip-chip bonded to sapphire substrate. (b) Quad VCSEL driver array before attachment. (c) Quad VCSEL driver array after attachment to VCSEL array

Several source-synchronous parallel optical interconnection designs in the application of multi-processors, chip-to-chip and board-to-board communication were reported in [42], [43], [44], [45], and [46], but the signaling schemes used in these systems were mainly based on SDR. In addition, the serialization and deserializaton digital circuits were located on a separate chip from the analog VCSEL driver and optical receiver circuits. To our knowledge, we are reporting the first demonstration of an integrated optical source-synchronous DDR transceiver.

The remainder of the section is organized as follows: Section 5-B introduces Peregrine Semiconductor's SOS process integration technology. Section V-C describes the mixed-signal IC design. Both analog transceiver circuits and digital logic circuitry that implements DDR serialization and deserializaton are described in detail. Section V-D describes the demonstration system, including PCB design and free-space optical link setup. Section V-E presents the electrical and optical testing results, and Section V-F is a conclusion section.

5.B. Integration Technology

The optical characteristics of the SOS process allow flip-chip integration of VCSELs and photodetectors directly onto the Ultra-Thin-Silicon (UTSi) substrate. As shown in Figure 5-2 (a), the active VCSEL apertures are bonded facedown on the SOS chip with the optical signals passing through the substrate. This allows low parasitic connections to the optoelectronic (OE) devices in a very simple physical package. Figure 5-2 (b) and (c) show an array of four VCSEL driver circuits at 250 um pitch before and after a quad VCSEL array is flip-chip bonded to it [47].

5.C. Mixed-Signal IC Design

5.C.1. IC Architecture

The transceiver IC consists of separate transmitter and receiver circuitry. Typically, the transmitter and receiver are integrated on the same chip, but due to chip area limitations, the VCSEL array and the photo-detector array were integrated on separate identical chips.

Figure 5-3 shows the functional block diagram of the ICs. The transmit section has two 8:1 serializers which convert a 16-bit parallel electrical CMOS input into two high speed serial data streams. The VCSEL drivers convert the serial data streams into optical signals, which carry the data across the free-space links. On the receive side the high-speed serial data streams are received and then deserialized back into a 16-bit parallel electrical output.

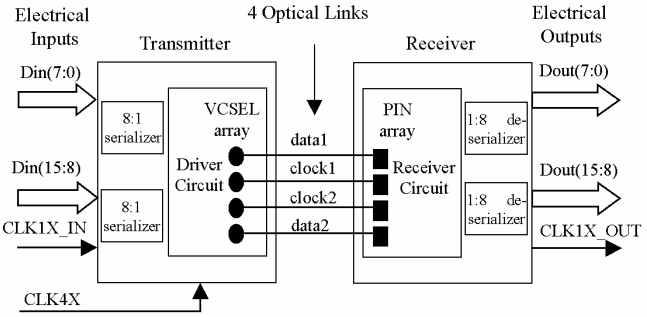


Figure 5-3. Functional block diagram of the DDR

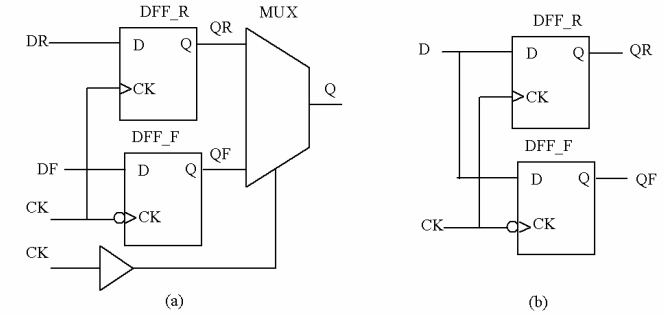


Figure 5-4. (a) Circuit diagram for DDR_MUX. (b) Circuit diagram

Quad VCSEL and photodetector arrays were flip-chip bonded at the center of the transceiver ICs, forming four optical links: two data channels and two clock channels. Since the clock is transmitted along with the data, clock recovery circuitry is not required on the receiver side and 8B/10B encoding is not required on the transmitter side.

5.C.2. Digital Circuit Design

The Serializer/Deserializer circuits were designed to convert wide slow parallel electrical signals to narrow fast serial data stream and vise versa. To implement the DDR scheme, a DDR multiplexer, (DDR_MUX), and a DDR demultiplexer, (DDR_DEMUX), each consisting of two flip-flops triggering on opposite edges of the clock signal, were incorporated into the Serializer and Deserializer circuit respectively. The DDR_MUX, shown in Figure 5-4(a), multiplexes the two input data into a serial stream at each edge of the clock whereas the DDR_DEMUX, shown in Figure 5-4(b), demultiplexes the serial input data into two data outputs at each edge of the clock. This section describes the digital circuit design of the DDR 8:1 serializer and the DDR 1:8 deserializer.

5.C.2.a. DDR Transmitter (8:1 Serializer)

Figure 5-5 shows the schematic diagram for one channel of the DDR serializer. It is composed of three major stages: the LOAD_GEN; two 4:1 parallel-in-serial-out (PISOs) converters, triggered at the rising and the falling edges of the clock respectively; and a DDR_MUX as the output stage of the serializer [48]. The inputs to the serializer are 8-bit parallel data and two clock signals, CLK1X (system clock, synchronous to the parallel data input) and CLK4X (four times as fast as the CLK1X). The output of the serializer is the serialized data stream that is triggered on both edges of CLK4X. A delayed version of CLK4X is sent along with the serial data as the accompanying clock. Special care was taken in the design to ensure that the serial data output and the delayed CLK4X output are exactly synchronous. The module LOAD_GEN generates the load pulses LOAD_RISE and LOAD_FALL, which are the inputs of PISO_RISE and PISO_FALL respectively. Figures V-6 (a) and V-6(b) are diagrams

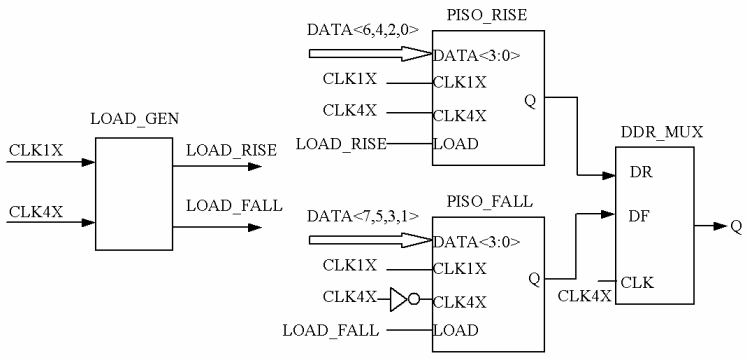


Figure 5-5. Schematic diagram for one channel of DDR 8:1 serializer

of the LOAD_RISE and LOAD_FALL generator circuits. Figure 5-6(c) shows a typical PISO (parallel in, serial out) circuit that was used in the design. A DDR_MUX, shown in Figure 5-4 (a), was used at the output stage to combine the output from both PISOs into a serial data stream at both edges of CLK4X. Figure 5-6(d) shows the transmitter output waveforms [48][49].

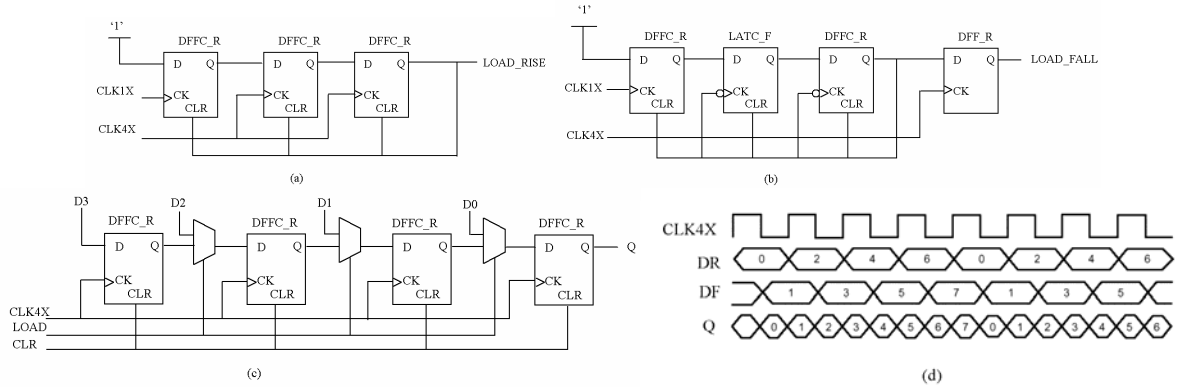


Figure 5-6. Schematic diagram for the components of DDR serializer (a) LOAD_RISE generator (b) LOAD_FALL generator c) Parallel-In-Serial-Out (PISO) (d) DDR_MUX output waveform

5.C.2.b. DDR Receiver (1: 8 Deserializer)

The DDR receiver accepts two serial data streams and their accompanying clock signals and generates a 16-bit parallel data stream. Figure 5-7 shows the gate-level schematic for one channel of the DDR receiver. It consists of three major stages: the Octal Data Rate Demultiplexer, the Clock Generator, and the Parallel Data Output Register at the end of the receiver [49].

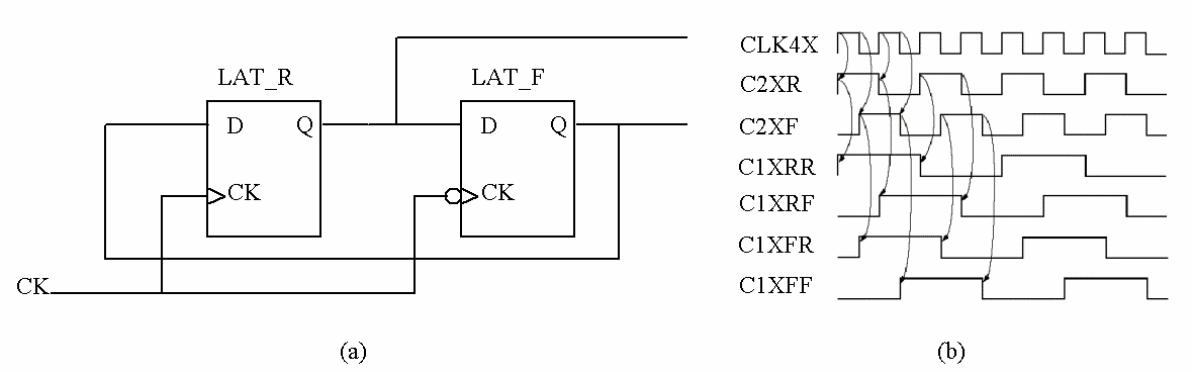


Figure 5-7. Schematic diagram for one channel of DDR 1:8 deserializer

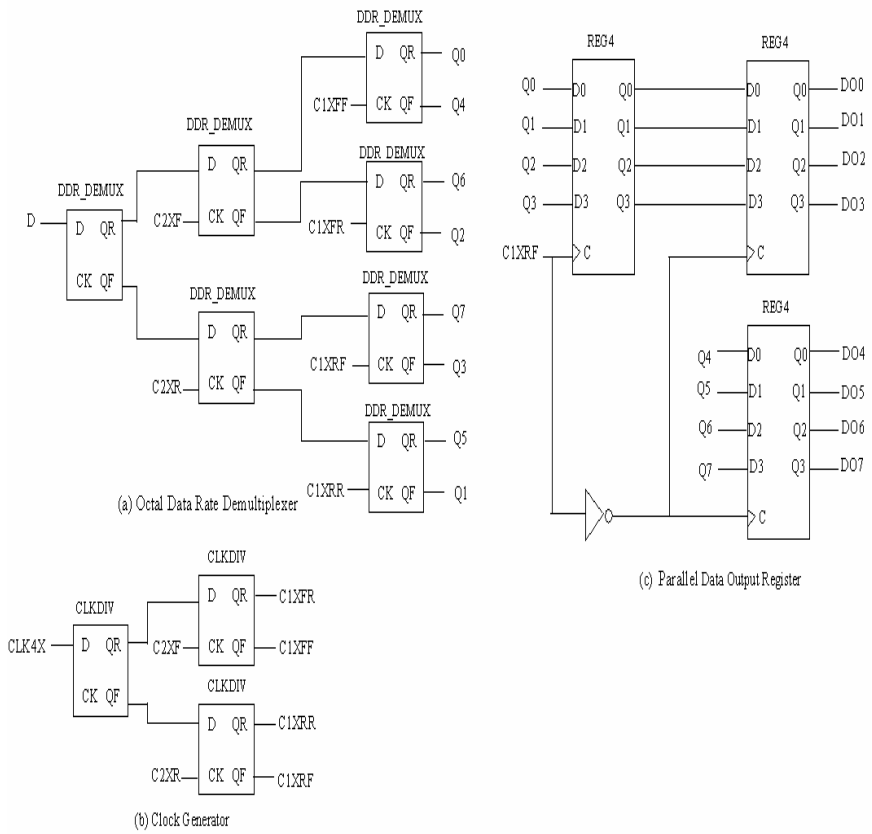


Figure 5-8. (a) Circuit diagram for CLKDIV. (b) Clock outputs waveform of the clock generator

Since we use a source-synchronous scheme (CLK4X is sent along with serial data stream), there is no need for PLL-based (Phase Locked Loop) clock recovery circuits on the receive side. Instead, half and quarter cycle phase clocks are generated from CLK4X by the Clock Generator for demultiplexing the received data. The Clock Generator, shown in Figure 5-7 (b), uses three clock dividers (CLKDIVs) to generate the multiple clock outputs. Figure 5-8(a) depicts the composition of the CLKDIV and Figure 5-8 (b) shows the waveform of the multiple clock outputs from the Clock Generator. As shown in Figure 5-8 (b), C2XR and C2XF are clock signals generated at the rising and falling edges of input CLK4X, which oscillate at half the rate of CLK4X. Similarly, C1XRR, C1XRF, C1XFR and C1XFF are clock signals generated at the rising and falling edges of C2XR and C2XF, and oscillate at a quarter of the rate of CLK4X. The Octal Data Rate Register, shown in Figure 5-7 (a) is a tree of DDR_DEMUX that uses the multi-phase clock outputs from the Clock Generator to deserialize the input data stream to an 8-bit parallel data output. The Parallel Data Output Register, shown in Figure 5-7 (c) samples the data from eight clock domains and latches it to a single clock domain. Its major component, REG4 is constructed of 4 bit parallel D flip-flops (registers) to latch the data in. The first four bits from the Octal Data Rate Demultiplexer (Q0-Q3) are captured by a REG4 on the rising edge of C1XFR and transferred to another REG4 on the falling edge. The second four bits (Q4-Q7) are captured by a REG4 on the falling edge of C1XFR. C1XFR is the system clock output aligned with the parallel data output.

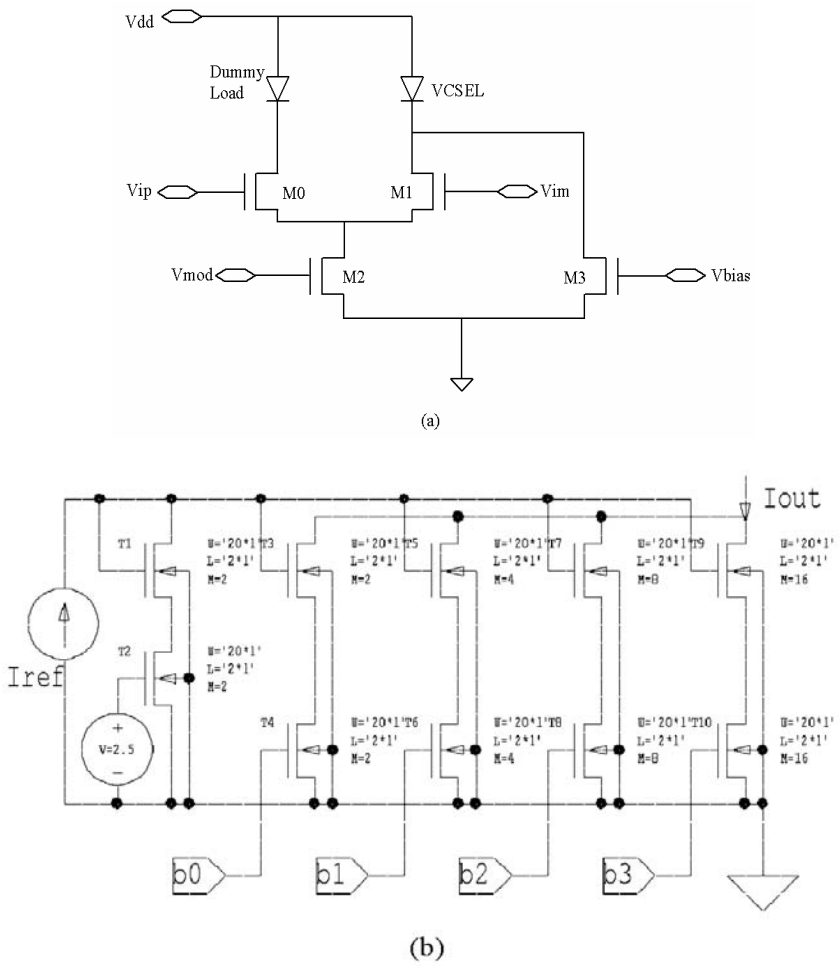


Figure 5-9. (a) A typical VCSEL driver circuit. (b) DAC setup for adjustable VCSEL driver output

5.C.3. Analog Circuit Design

5.C.3.a. VCSEL Driver Circuit Design

A typical VCSEL driver circuit uses a differential current-steering topology as shown in Figure 5-9(a) [50], [51], [52]. The VCSEL is connected to the right-side output and a dummy load is connected to the left-side output of the amplifier. Constant bias current is supplied by transistor M3 to ensure that the VCSEL is always operating above its threshold current while M0 and M1 are differentially driven to switch modulation current through the VCSEL.

In our driver circuits, we replaced the analog current sources M2 and M3 with digitally programmable current-mode digital-to-analog converters (DACs), shown in Figure 5-9(b). This enables us to dynamically adjust the optical output power. Each VCSEL has its own individual bias and modulation current setting for greater flexibility. The digital settings are stored in on-chip digital registers.

5.C.3.b. Photo-Detector Receiver Circuit Design

The photo-detector receiver circuitry is composed of a transimpedance amplifier (TIA), an RC filter, a decision circuit, two post-amplifier stages and a ToCMOS stage as shown in Figure 5-10 (a). The TIA converts photo-detector current into a voltage signal. The RC filter is used to find the average value of the input signal. The decision and post-amplifier stages are differential amplifiers that amplify the signal to digital current mode level (CML). The ToCMOS stage converts current mode level signal to CMOS signal for digital processing.

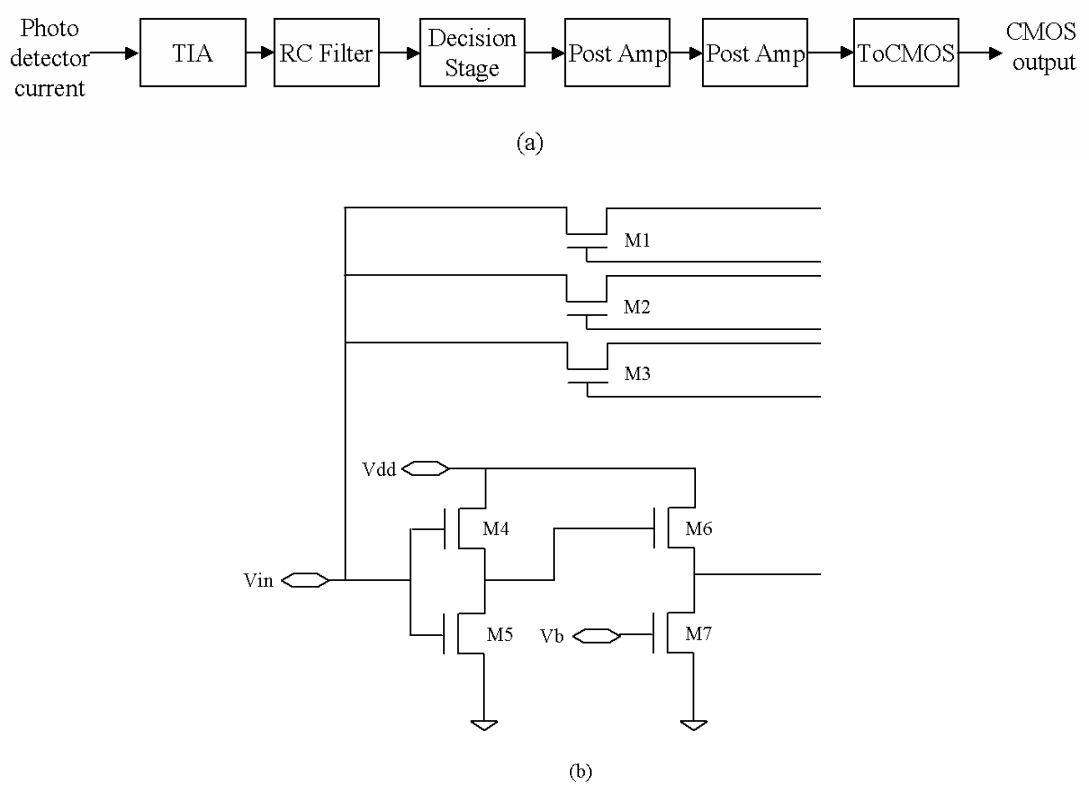


Figure 5-10. (a) A typical VCSEL driver circuit. (b) DAC schematic for adjustable

Figure 5-10(b) shows the schematic of the TIA design. Three n-type MOSFETs with various sizes were used in the feedback paths as feedback resistors. Each of these MOSFETS is digitally controllable, allowing for dynamic adjustment of the gain of the TIA.

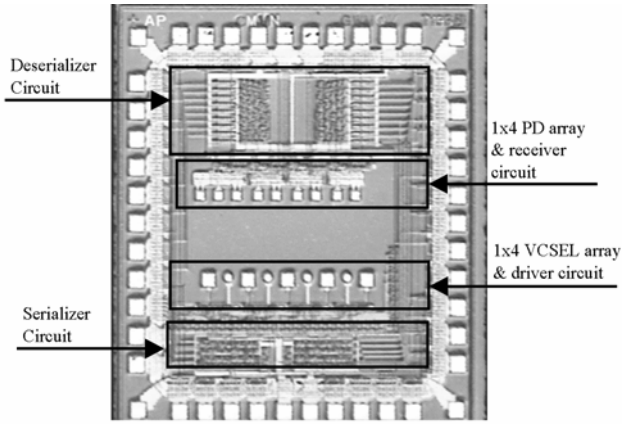


Figure 5-11. Microphotograph of the transceiver IC

5.C.4. Chip Layout

Figure V-11 shows the microphotograph of the DDR transceiver IC, fabricated using the SOS 0.5um UTSi CMOS process. The chip dimensions are 2.3mm x 2.7mm and it has 44 VCSEL driver power/perimeter I/Os. Due to pad limitations, bi-directional I/O pads were used for the 16 electrical I/Os with one extra pad to configure them as either input pads in the transmitter or output pads in the receiver.

5.D. System Integration

5.D.1. OE Device Attachment

The VCSEL array operates at 850 nm with threshold current in the range of 1.5 ~ 2.0 mA. The differential resistance at 4 ~ 8 mA is 50 ohms and slope efficiency is 0.45 mW/mA. The die size of the VCSEL array is of 1.2 mm x 0.45 mm with 250 um pitch between each channel [16]. The Gallium Arsenide (GaAs) PIN Photodiode array operates with a responsivity of 0.5 A/W at 850 nm. The size of the array is 1.055 mm x 0.45 mm and the pitch between devices is 250 um [54].

Unlike backside-emitting VCSELs which necessitate the removal of the GaAs substrate [51][55], the sapphire substrate allows VCSEL and photodiode arrays to be flip-chip bonded face down to the center of the transceiver IC with the optical signals passing though the substrate.

5.D.2. Test-Bed System

To test the operation of the IC we built a PCB main board and two chip-carrier boards, one for the transmitter and one for the receiver. The transceiver ICs with OE (opto-electronic) arrays attached were wire bonded to the carrier boards and sealed with epoxy. A small rectangular section of the carrier board under the IC was removed to allow optical access to the OE arrays. The carrier boards were then mounted to the main board with high-speed surface mount connectors. The distance between the carrier boards on the main board is about 76.2 mm. Figure 5-12 shows a schematic of the test-bed system.

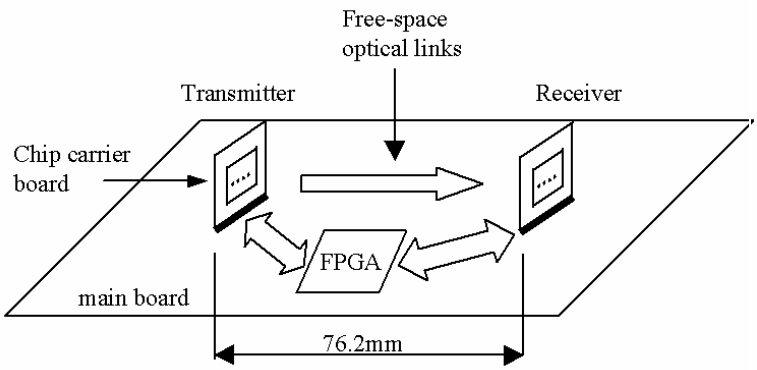


Figure 5-12. Schematic of the test-bed system

The main board is a 7 x 8-inch eight-layer FR4 board. A Xilinx Virtex FPGA was placed in the center of the board and was programmed to control the test of the transceiver ICs. The Virtex FPGA has a built-in delay locked loop (DLL), which we used to generate CLK1X_IN and CLK4X (see Figure 5-3) in perfect phase synchronization for the transmitter. Figure 5-13 shows the complete test-bed system assembly.

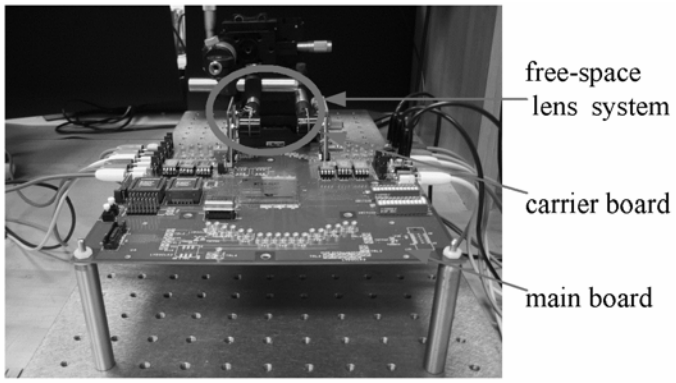


Figure 5-13. Free-space optical demonstration system setup

5.D.3. Free-Space Optical Link Setup

We interconnected the chips using the free-space setup shown in Figure 5-14. Starting with 15 high-resolution f-1.2 seven-element Universe Kogaku Double-Gauss lenses, we selected two lenses that were well matched. Placed at one focal length from the VCSEL array, the first lens collimated the VCSEL beams, while the second lens re-imaged them onto the detector array. Because the lenses were well matched the magnification error between the imaged VCSEL array and the photodetector array was minimal. Since the optical system was approximately paraxial, the optical path length (OPL) for each channel did not vary significantly over the array. As a result of this fact, the skew between the different channels was very slight. The horizontal pitch between OE devices is 250 um and the off-axis vertical distance of the VCSEL array is .385 mm and -.385 mm for the detector array. We used an approximate lens model in an optical design program to estimate the OPLs for the four channels. The results are tabulated in Table 5-1 below. The transmission latency for each of the four channels resulting from these estimated OPLs are similarly shown. The maximum transmission latency difference between channels is less than .01 picoseconds. The optical system was designed to permit a reflective neutral density filter to be placed between the lenses in order to allow simultaneous oscilloscope observation of transmitter and receiver signals, which also allowed visual verification of the performance and alignment of the lenses with a Charge Coupled Device (CCD) camera. Using a reflective neutral density filter with an optical density of .03, a portion of the signal was split off and coupled with a fiber-coupled detector while the rest impinged upon the detectors. It was not possible to achieve this kind of observation using an infrared sensitive camera, because there was too much reflection from the lens surfaces as well as problematic CCD blooming. To overcome this issue we used a color camera that was not sensitive to infrared. Using the color camera allowed us to observe the incident beam spots without the problem of CCD blooming because, although the VCSELs peak in the infrared, they have a small percentage visible spectral content in the red. Figure 5-15 below shows a view of the IC with OE detector arrays as seen by the camera.

5.E. Experimental Results

We performed both electrical and optical experimental measurements on the IC and the optical link between the transmitter and receiver. Test results show that both the IC and the optical links are fully operational at a data rate up to 500 Mbps. Table 5-II below lists the operating power consumption per channel at 500 Mbps data rate.

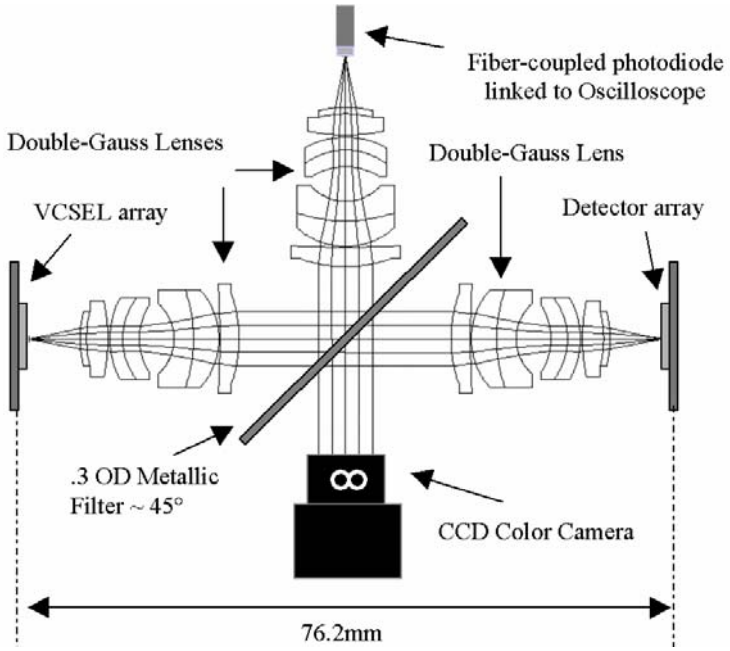


Figure 5-14. Schematic of the optical lens system design

Table 5-1. Optical path length (OPL) and transmission latency of each of the four optical channels

VCSEL #	Vertical (mm)	Horizontal (mm)	OPL (mm)	Latency (ps)
1	0.385	-0.375	91.701	305.670
2	0.385	-0.125	91.703	305.677
3	0.385	0.125	91.703	305.677
4	0.385	0.375	91.701	305.670
			Average OPL 91.702	Average Latency 305.673

We measured DC characteristics of the transimpedance amplifier (TIA). Figure 5-16 shows that the DC characteristic curve matches the simulation results almost exactly. The same TIA output was monitored during the process of aligning the optical elements to fine-tune the alignment of the lenses in the system.

We then performed electrical testing on the IC without OE devices attached to it. The IC was wire bonded to a universal test board where we measured the electrical IO outputs as well as the output on the VCSEL pads. Figure 5-17 (a) is a snapshot of the stimulus to the transmitter generated from Logical Analyzer and Figure 5-17 (b) shows the corresponding voltage output measured on the VCSEL pads. The measurement was done using a 50-ohm surface-mount resistor to emulate the VCSEL device. The output was verified to be exactly the serialized version of the parallel electrical input to the IC, triggering on both rising and falling edges of the CLK4X signal.

By using the test-bed system shown in Figure 5-13 and Figure 5-14, we performed a complete test on the whole system from the transmitter to the receiver through free-space optical links. As mentioned above, a portion of the optical signal was split off and coupled with a fiber-coupled photodetector, which was monitored by an oscilloscope. Since only a small portion of the optical signal was used in this measurement, the receiver was simultaneously able to continue receiving the optically transmitted data (see Figure 5-14). Figure 5-18 shows a captured optical data waveform at 500 Mbps. The optical data was verified to be the same as the deserialized output displayed by the receiver.

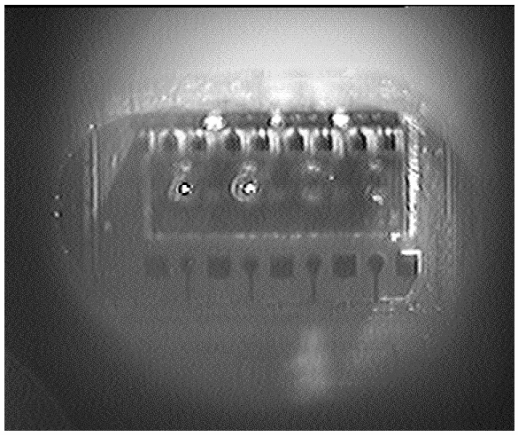


Figure 5-15. A view of IC with OE arrays attached as seen by the camera

Table 5-2. Power consumption of analog and digital circuits

	Current (I)	Voltage	Power
VCSEL driver circuit	7.2 mA (5.4mA+1.8mA)	3.3V	24mW
Receiver circuit	10 mA	3.3V	33mW
Digital circuits (DDR SerDes)	7 mA	3.3V	22mW

The bias current and modulation current of the VCSEL driver were set at 1.8 mA and 5.4 mA respectively. Eye-diagram measurements were performed on the transmitter optical serial output. Pseudo random data (2^8-1) in 8-bit parallel form was generated for each channel inside the FPGA using Linear-Feed-Back-Shift-Registers (LFSR). The random data was sent to the transmitter chip, which generated the serialized data stream through DDR serializer after which optical signals were generated. The optical signal was measured using the fiber-coupled detector mentioned earlier. Figure 5-19 shows the measured eye diagrams at data rate of 160 Mbps and 500 Mbps respectively. As can be seen from the figure, the eye diagram is very open, indicating low bit error rate, even at 500 Mbps.

5.F. DDR Parallel Optical Interconnect Summary

A two-channel source-synchronous parallel optical transceiver IC using double data rate (DDR) clocking has been designed and implemented. This design demonstrated advanced methods of improving data communication with reduced power operation, and is therefore directly applicable to the PCA problem domain, where power consumption at the communication link is important because overall power consumption and heat dissipation are a concern for high-performance processors. With the DDR scheme, the system can immediately achieve twice the SDR bandwidth with any given clock speed. A free-space optical demonstration system was built to test the functionality of the chip. Testing results have shown the chip and the link are fully operational at 1Gbps aggregate data rate with 500 Mbps per channel across a free-space link. The success of this first effort shows that novel signaling schemes may promote the use of optics in very short-range, highly-dense, inside-the-box applications, which were once the domain of only wires and vias.

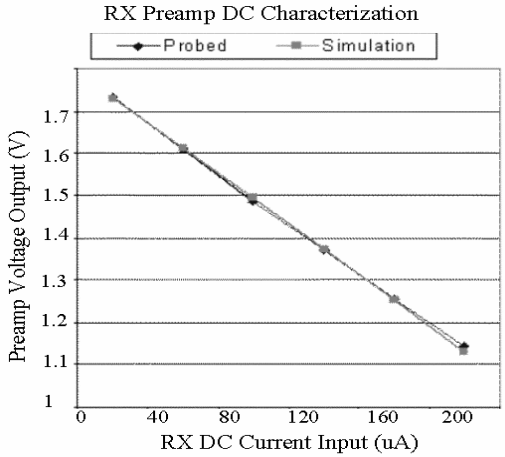
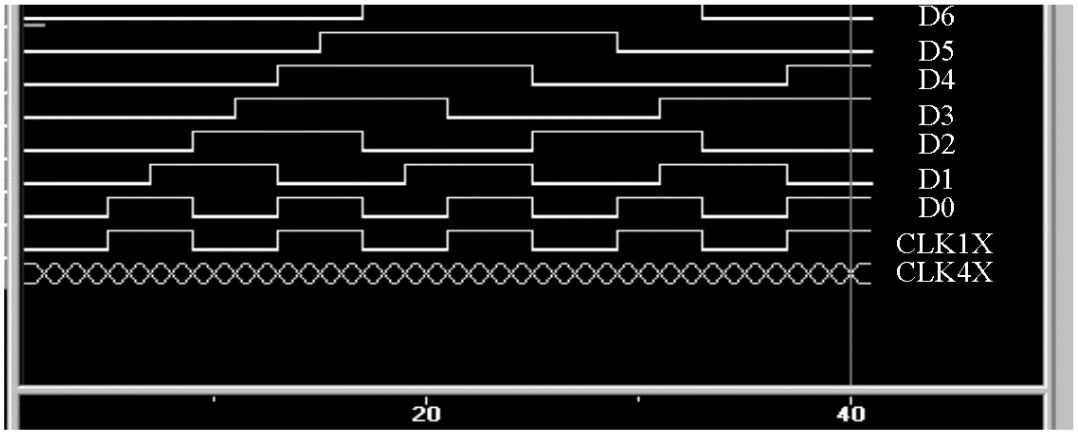
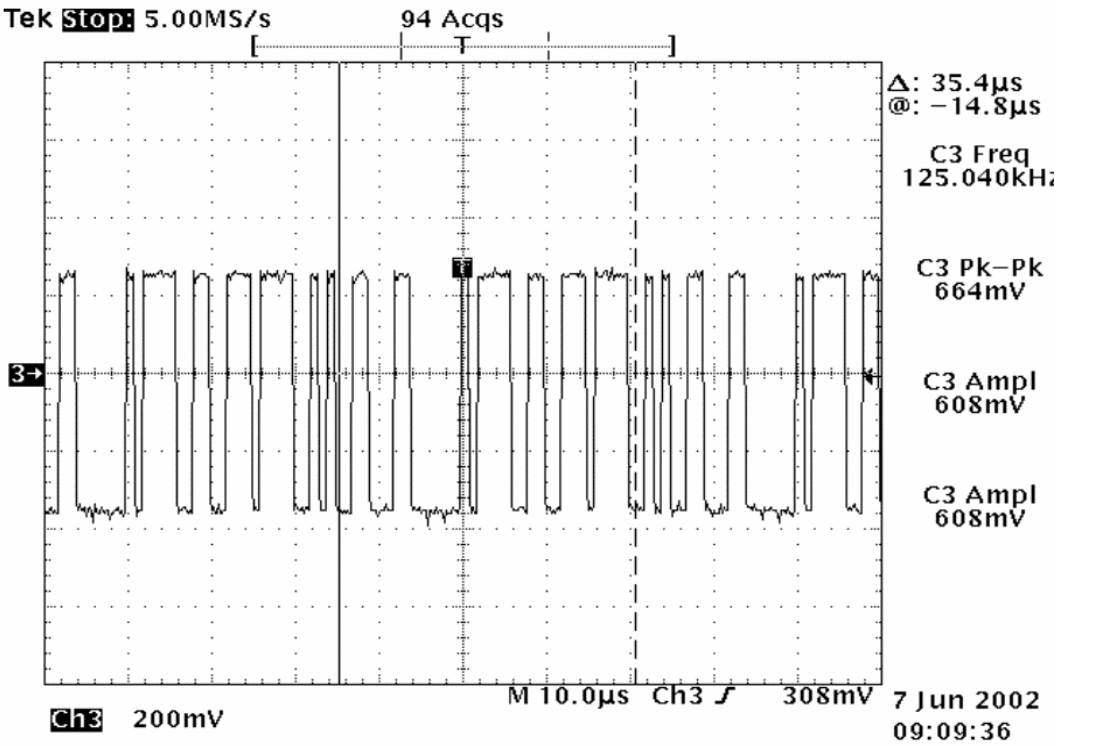


Figure 5-16. DC test output vs. simulation output of the TIA



(a)



(b)

Figure 5-17. (a) Snapshot of the stimulus from logical analyzer. (b) Electrical data output measured on the VCSEL pads by emulating circuits

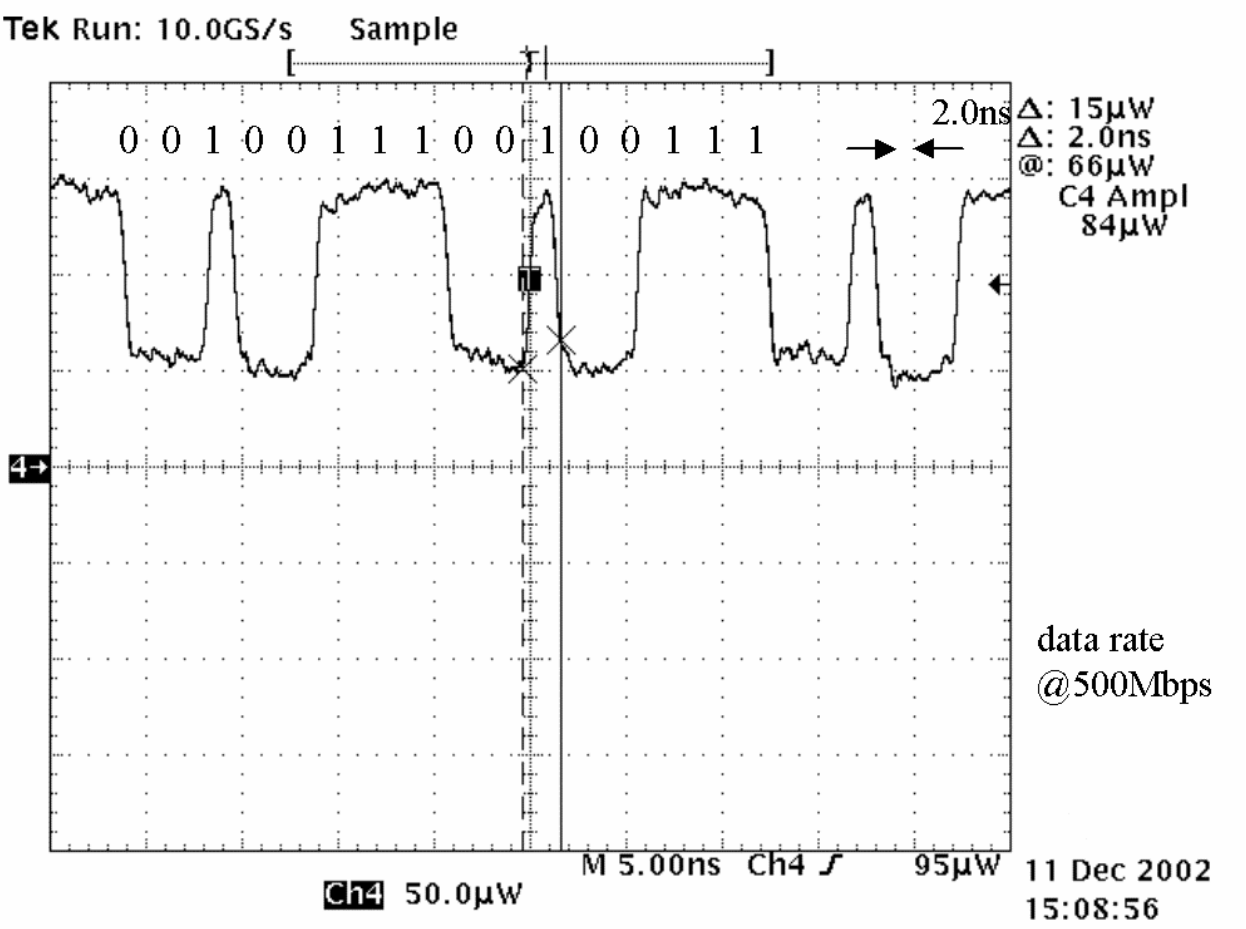


Figure 5-18. Optical serial data stream (00100111) captured by oscilloscope through optical probe. The minimum pulse width is 2ns as shown in the figure with data rate of 500Mbps per channel. Since DDR clock scheme was used, CLK4X was at 250 MHz, half the data rate

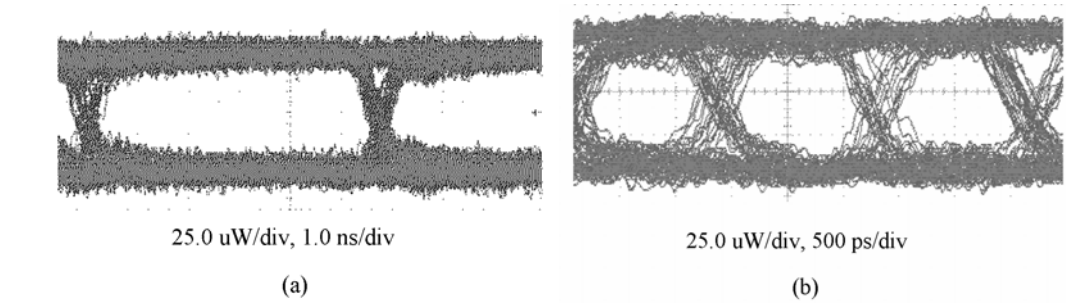


Figure 5-19. Measured eye diagram (a) At 160 Mbps. (b) At 500 Mbps

6. PERFORMANCE-BASED POWER OPTIMIZATION FOR PACKAGING-FRIENDLY PARALLEL OPTICAL TRANSCEIVERS

6.A. Power Optimization Introduction

High-density, small spot size, and high power consumption have imposed both a technical and an economic challenge on optoelectronic (OE) packaging and must be addressed for OE devices to be integrated into PCA systems. Device non-uniformity and dynamic operating environments cause device performance variations, especially in applications where large OE arrays are employed. We describe our work to explore the opportunities in transceiver circuit design to relax optical alignment requirements, provide efficient thermal management, and allow dynamic variation within the optical system. Our approach focuses on a built-in power negotiation algorithm, which is developed to dynamically optimize the power consumption of optical links based on the bit error rate (BER) of each optical link. This algorithm can be executed as the system is powered up or during normal system operation if the link is idle or a change has been made. It converges to an optimal setting for each VCSEL that has the minimum power consumption for a given target BER. The intelligence of the dynamic power optimization provides the system with the capability of compensating for the potential variations in the system and therefore, relaxing the packaging requirements. We implemented this algorithm in a 0.5μm CMOS silicon-on-insulator (SOI) chipset and developed parallel optical transceivers based on VCSELs and photodetectors to demonstrate the benefit of our approach. We built a field programmable gate arrays (FPGA) based test bed where both fiber optic and free-space optical interconnects are used for communications between two chips. Test results show that the algorithm is able to find the optimum power setting for all VCSELs despite varied light attenuation in optical path.

High density and very-large-scale optoelectronic integration (OE-VLSI) allows great flexibility in increasing aggregate bandwidth. However, increasing speed and tight integration lead to increased power consumption per unit volume. High-density, small spot size and high power consumption have imposed both a technical and an economic challenge on OE packaging where currently packaging costs are dominant in such product development [62]. To illustrate, consider that III-V optoelectronic devices (made of elements from columns III and V of the periodic table, e.g. GaAs) typically have small and asymmetrical spot sizes that make coupling to single-mode fibers very inefficient and alignment sensitive. Furthermore, high power consumption and the lack of efficient cooling lead to overheating of components resulting in performance and lifetime degradation. Temperature fluctuation and manufacturing process variation introduces deviations in threshold current, slope efficiency and series resistance of the VCSEL, which eventually introduces jitter to the optical signal and degrades system performance [63]. These factors lead to a need to carefully develop sophisticated packaging strategies for systems using optoelectronic devices.

A lot of work has been done to develop affordable, compact and reliable OE packaging. This work is mainly focused on producing new and innovative optoelectronic devices, developing new alignment tolerant structures and subsystems, improving optical coupling efficiency, increasing the compatibility with existing integration processes to enable large-scale integration, and automating the alignment and assembly process to lower the packaging cost [61][62][64-66].

With the goal of simpler OE packaging, opportunities in transceiver circuit design to relax optical alignment requirements, provide efficient thermal management, and allow dynamic variation within the optical system have been investigated. The approach is to develop a power negotiation algorithm for optical transceivers that can dynamically find the optimal power setting to achieve a target performance. Digitally tunable drivers and receivers are uses to offer individual control for each OE device and to allow adaptability to varying light power and device parameter variation. Theoretical analysis of optical communication channels reveals the dependence of channel BER on the optical power. Based on this dependence, a power negotiation algorithm was developed to adjust the power setting of optical links based on the bit error rate of each optical link with the system setup as shown in Figure 6-1. This algorithm can be executed when the system is first powered on or during normal system operation if the link is idle or a change has been made. It converges to an optimal setting for each VCSEL that has the minimum power consumption for a given target BER.

To demonstrate the benefit of the approach, this algorithm was implemented in a 0.5μm CMOS SOI chip and integrated 1x4 VCSEL and photodetector arrays using flip-chip bonding. An FPGA-based test bed was built in which both fiber optic and free-space optical interconnects are used to establish communications between two chips. The algorithm was demonstrated to find the optimum modulation and bias settings for VCSELs for a target BER despite changes in the system- operating environment such as varying optical loss. The rest of this section will start with the analysis of how the bit error rate depends on operating conditions in section 6-B. We present the details of the algorithm and its implementation in section 6-C. The custom transceiver design and chip layout will be introduced in section 6-D. The integration of optical demonstration system will be given in section 6-E and the test result will be presented in section 6-F.

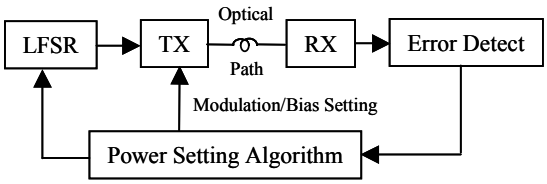


Figure 6-1. Power negotiation algorithm block diagram

6.B. Theoretical Background

6.B.1 Equations

For a single-channel on-off keyed intensity-modulated link where the noise is assumed to be Gaussian distributed, the BER with the optimum decision threshold can be given by

$$P_e = 0.5 erfc\left[\frac{Q}{\sqrt{2}}\right] \qquad \text{[VI-1]}$$

where $Q=(I_1-I_0)/(\sigma_0+\sigma_1)$ is the signal-to-noise ratio (SNR) at the decision circuit. I_1 and I_0 are the induced photocurrent from the detector and σ_1 and σ_0 are the rms noise when bit 1 and bit 0 are received respectively. In (VI-1), *erfc* stands for the complementary error function, defined as

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-u^2} du \qquad \text{[VI-2]}$$

Since the induced photocurrent (I_1 or I_0) is proportional to the optical power incident on the detector during the data cycle, and the link optical power is determined by the driving current for the VCSEL (I_{on} for data 1 and I_b for data 0) with direct intensity modulation scheme, equation (VI-1) can be and simplified as

$$P_e = 0.5 erfc\left[\frac{\gamma(I_{on} - I_b)}{\sqrt{2}}\right] \qquad \text{[VI-3]}$$

where γ is the scaling factor accounting for electrical-to-optical (E-O) and optical-to-electrical (O-E) conversion, channel attenuation and noise for a given system configuration.

Equation (VI-3) is true only for a VCSELs biased above its threshold current. However, when a VCSEL is biased below its threshold current (I_{th}), a significant turn-on delay occurs at the rising edge of optical pulse [68]. This delay is a result of the time it takes for the photon population to build up in the laser cavity after the carrier density has been decaying during its "off" period: i.e. when it is biased below threshold. The turn-on delay varies from pulse to pulse depending on the data rate and the total off period prior to the pulse, which results in pattern-

$$P_e = 0.5 erfc\left[\frac{\gamma(I_{on} - I_{th})Cos(\pi B \tau F)}{\sqrt{2}}\right] \qquad \text{[VI-4]}$$

dependent jitter. An approximate form for the worst-case error rate below threshold is [69][70]

$$F = Ln\left[\frac{I_{on} - I_b}{I_{on} - I_{th}}\right] \qquad \text{[VI-5]}$$

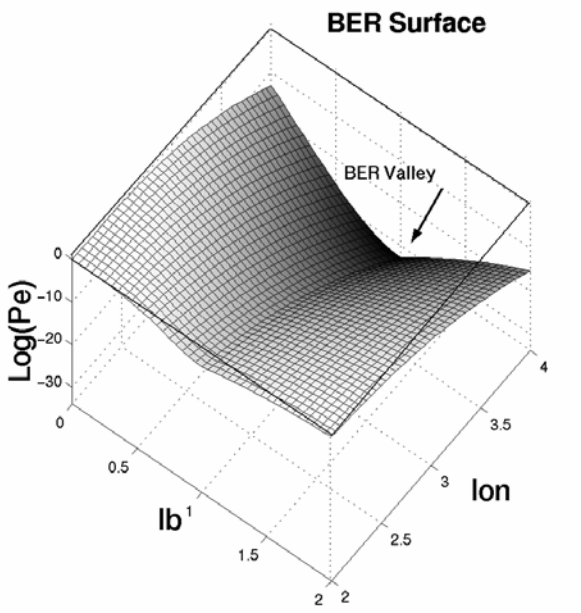


Figure 6-2. Bit error rate surface as a function of transmitter power setting (Ion and Ib) shows the BER "valley" where the minimum BER exists for an optical link

where B is the bit rate, τ is the carrier recombination time and while expression for the error rate is approximate, it models the fact that for below threshold conditions, the error rate increase due to pattern dependent jitter. Therefore, complete representation of the BER as function of transmitter power settings can be determined by combining (3) for biasing above threshold current with (VI-4) for biasing below threshold current. For $B\tau=1$, $I_{th}=1$, and $\gamma=4$, we plot (VI-3) and (VI-4) in the 3-dimentional graph which is shown in Figure 6-2.

Figure 6-2 shows how the BER changes with varied power settings. The X and Y axes represent values of I_{on} and I_b, while the Z axis varies with $\text{Log}(P_e)$. With a fixed I_{on}, the minimum BER appears at the point when I_b is equal to I_{th}. Varying I_b from I_{th} in either direction will result in degrading BER. On the other hand, for a given I_b, BER is decreasing with increased I_{on}. In other words, a BER "valley" exists parallel with the I_{on} axis at $I_b=I_{th}$, where the lowest BER is located. Therefore, this BER "valley" indicates there is an optimal power setting that results in minimum BER.

To further illustrate the dependence of BER on power setting, we plot the BER contours with BER ranging between 10^{-6} and 10^{-15} in Figure 6-3 (a). Each curve corresponds to a certain value of BER and the X and Y coordinates of any point on the curve represents the power setting (I_b and I_{on}) that results in that BER. For a given BER, I_{on} gets the minimum value at threshold biasing. The average power consumption of the laser can be approximated by $P_{elec} \approx 0.5V_{on}(I_{on}+I_b)$ where V_{on} is the turn-on voltage of the laser. To minimize its power consumption for a certain BER value, we only need to find out the power setting along the BER contour that results in the minimum of sum of I_b and I_{on}. Figure 6-3 (b) shows a comparison between two plots with $B\tau=1$ and $B\tau=2$. The value of $B\tau$ has no impact on the curves for above-threshold biasing and it only affects the slope of curves for below-threshold biasing. As $B\tau$ increases, the slope becomes greater. Therefore, for a reasonably large $B\tau$, because I_{on} increases on both sides of threshold-biasing rapidly, it can be seen that the power setting at the BER valley leads to the minimum sum of I_b and I_{on}, and thus, the minimum power dissipation. However, at small $B\tau$, typically as $B\tau$ drops below 1.2, the minimum power setting will no longer be at threshold biasing point. Instead, zero-bias is the lowest-power solution [69]. We assume that the value of $B\tau$ is greater than 1.2 unless the exception is explicitly stated. This assumption is justified because for values of τ on the order of 1 ns, this corresponds to a signaling rate of approximately 1.2 Gb/s. This rate is the approximate signaling rate of current Gb/s physical layer standards such as Fiber Channel and Gigabit Ethernet that use 8B/10B line coding. Thus the use of the BER valley will produce minimum power for existing high speed interconnect standards.

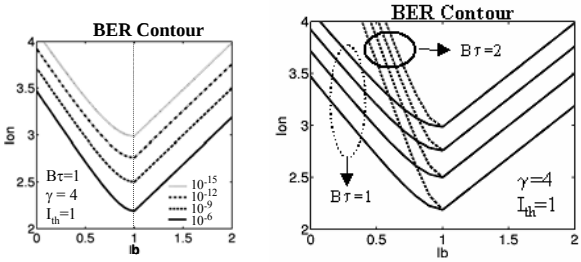


Figure 6-3. Bit error rate contours: (a) Showing the BER dependency on Ion and Ib and BER valley for optimum power setting (b) Showing how the value of Bt impacts the BER contour for below-threshold biasing

In summary, for any given target BER, the power setting in the BER valley results in minimum power consumption. Therefore, if we can develop an algorithm to locate the valley and use the optimum setting for the transmitter, it will lead to

the lowest power dissipation for this link. In addition, the algorithm needs to be able to recover the optimum power setting by reexamining the BER information when link performance is impacted due to the changed parameters, such as Bτ, γ, and I_{th}.

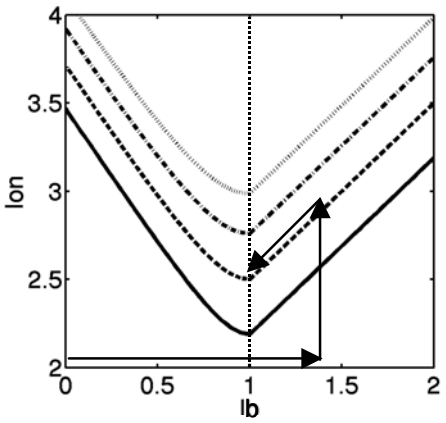


Figure 6-4. Decision algorithm shows the steps to find the optimum power setting

6.C. Power Negotiation Algorithm

The complete power negotiation algorithm is composed of two parts: a bit error rate tester (BERT) and a decision algorithm. The BERT provides the BER information of optical link. The decision algorithm adjusts the power (I_{on} and I_b) based on the channel BER information.

The goal of the decision algorithm is to find the optimum power setting prior to the real data transmission. Following the steps below will lead to the optimal settings as illustrated in Figure 6-4.

1) Set the bias current I_b above the threshold current suggested by the device manufacturer and start I_{on} with low end.
2) Fix I_b and keep increasing I_{on} until the target error rate is achieved.
3) Decrease both I_b and I_{on} together while the BER is less than the target BER.

The algorithm converges to the power setting corresponding to BER valley for Bτ>1.2. In the case of lower data rate systems when Bτ<1.2, it may result in below threshold biasing as the final power setting and the average power consumption may be slightly higher than the one biased at zero current.

Figure 6-5 shows the flow chart of the negotiation algorithm implementation. It starts with the transmitter sending a PRBS (pseudo-random binary sequence) pattern over the optical link to a receiver. The receiver checks incoming data and records the number of errors. Once the data transmission is done, the error information is sent back to the transmitter and the decision algorithm makes the proper adjustments to the power setting for the VCSEL. Therefore, each iteration involves one BER collection and one power setting adjustment. This iteration repeats until the optimum setting has been found.

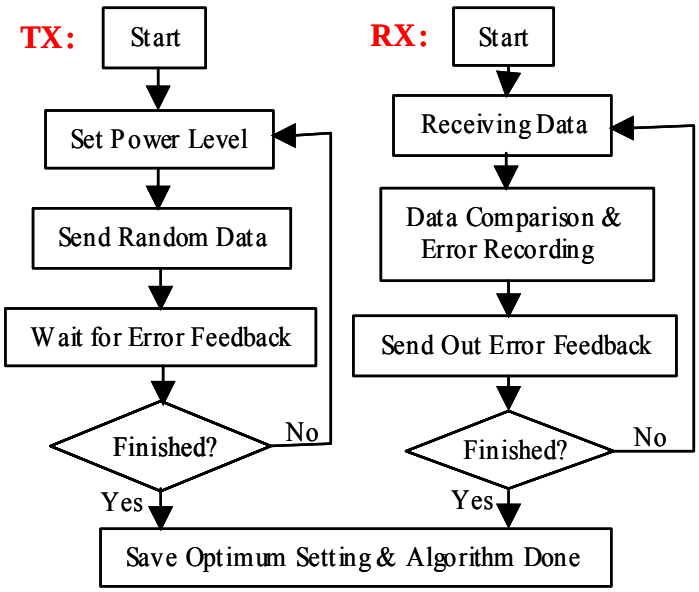


Figure 6-5. Power negotiation algorithm design flow

This negotiation algorithm requires a bi-directional path: a forward data path and an error feedback path. However, the error feedback does not require high-speed links and one feedback link can be shared among multiple receivers. In our case, an electrical feedback path is used to pass the error information.

6.D Hardware Implementation

We have fabricated a CMOS chip and built an FPGA-based demonstration system to verify this algorithm. Four uni-directional optical links have been established between two chips. To provide maximum flexibility in algorithm development and testing, only a part of the algorithm logic has been integrated into the chip, which includes the PRBS generator on the transmitter side and the bit error rate tester (BERT) on the receiver side. All decision algorithm and other associated control logic has been implemented in the FPGA.

6.D.1. CMOS IC Design

6.D.1.a. Chip Architecture

The CMOS IC chip we fabricated contains four optical transceiver designs, core algorithm digital logic and the interface with OE devices. The architecture of the chip is shown in Figure 6-6.

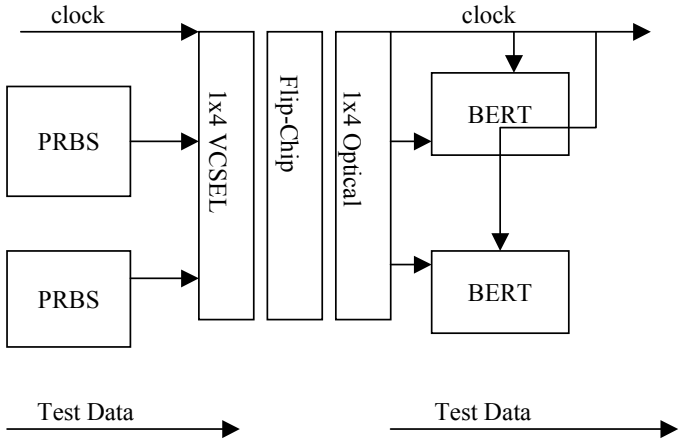


Figure6-6. CMOS chip architecture overview

On the transmission side, a system clock that is brought in electrically is used to generate the PRBS and is transmitted optically at the same time. On the receiving side, the recovered clock is used for data recovery and data comparison in the BERT. A special character, K28.5, is used for synchronization purposes. Due to limited chip area, only one OE array–either VCSEL or photodetector-can be integrated with one CMOS chip at a time.

6.D.1.b. VCSEL Driver Design

Since the light output of a VCSEL is linearly proportional to its driving current in normal operating range, the goal of the driver design is to provide digitally tunable modulation and bias current for VCSELs. Four-bit digital-to-analog current converters (DACs) are used which allows digital control over the transmitter. Therefore, control settings can be loaded and stored in register-based or RAM-based digital storage cells. Each transmitter has two current DACs, one for modulation current and one for bias current control. This, in turn, provides independent settings for each VCSEL and allows great flexibility in compensating for device non-uniformity and device parameter variation across a large OE array.

Figure 6-7 shows the schematic of a current DAC with 4-bit digital control inputs b0-b3. It is based on a current mirror structure and the transistors in each mirrored branch are sized proportionally so that each bit has different significance of current control. The reference current input, Iref, is the step size for current tuning and can be adjusted externally by the user to suit the requirements of different systems.

Figure 6-8 shows a comparison of the current output of the DAC between simulation and probed data. Close

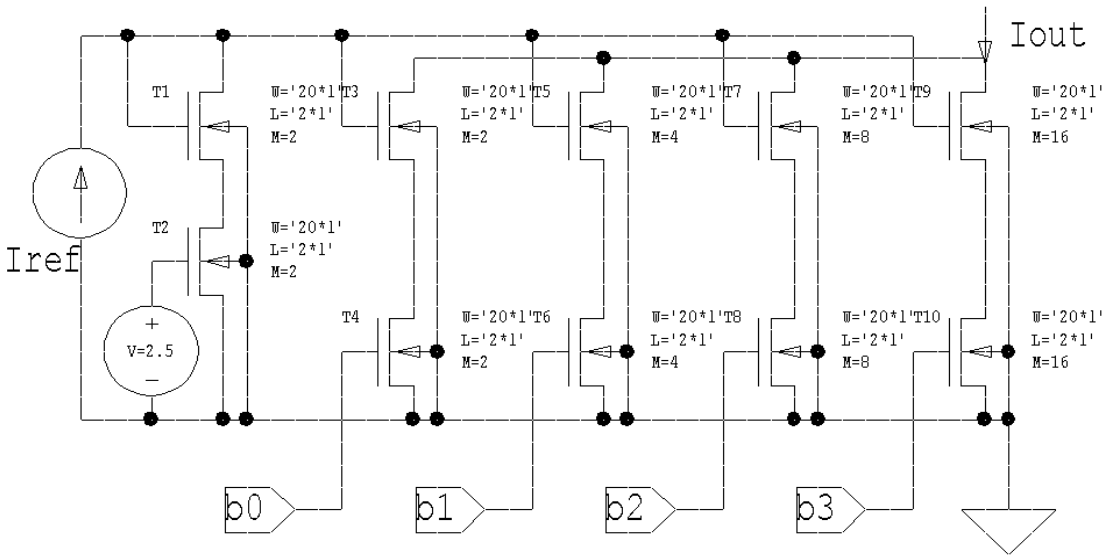


Figure 6-7. Digital-to-analog current converter

correlation between the two can be observed. Each curve illustrates the linearity of current changing while sweeping the digital settings from "0000" to "1111" with step size of I_{ref}, which, in the end, ensures the precise control over the driving current for VCSELs.

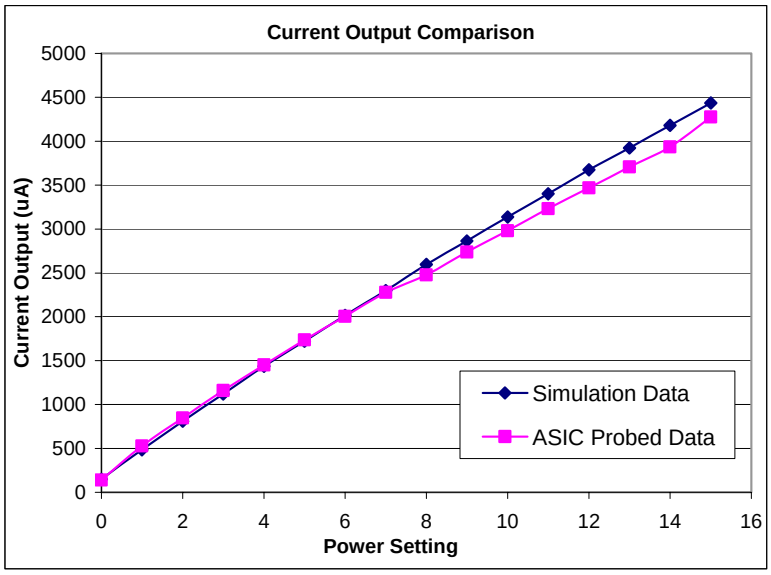


Figure 6-8. DAC output current comparison

6.D.1.c. Optical Receiver Design

The receiver is designed in such a way that it complements the operation of the transmitter. It has a large dynamic range with the ability to trade between gain and bandwidth during operation. Like the transmitter, the receiver also has digital control of its gain settings. The current receiver design consumes about 30 mW in average while operating.

Figure 6-9 shows the block diagram of the receiver design. It is composed of a transimpedance amplifier, three stages of differential amplifiers and a CMOS conversion stage. The transimpedance amplifier converts the current input to voltage signals and feeds it into the first

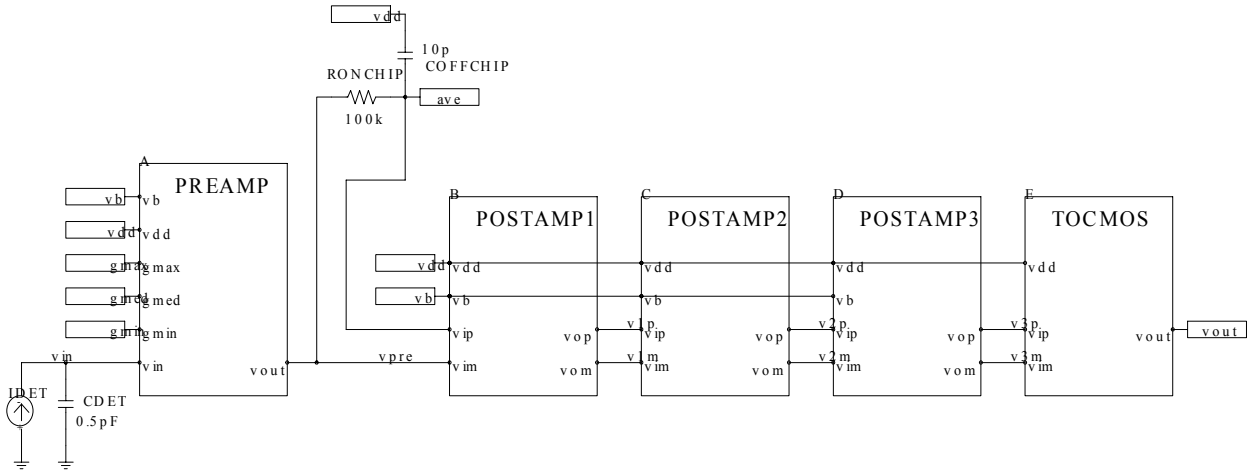


Figure 6-9. Optical receiver block diagram

differential amplifier. An external DC reference voltage can be applied to the other input of differential amplifier or an off-chip capacitor can be used to generate the reference voltage automatically. The CMOS conversion stage converts the amplified differential signal to a single-ended CMOS signal.

Figure 6-10 shows the schematics of pre-amplifier of receiver design. Three digital inputs (gmin, gmed, and gmax) are the digital gain control inputs, which give 1K, 2.5K and 10K ohms transimpedance gain respectively when enabled. A close matching is observed between simulation and test results for the preamplifier DC performance as shown in Figure 6-11.

6.D.1.d. Chip Layout

Our chip was designed and fabricated by using the Peregrine 0.5 μm ultra-thin silicon on sapphire (UTSi/SOS) process. UTSi is a newly developed SOI technology where devices are made in a very thin silicon layer (100nm) on top of an insulating sapphire substrate. Its advantages include: low power consumption, reduced parasitic capacitance, minimum crosstalk and an optically transparent substrate. The sapphire substrate is also a good thermal conductor and the thermal expansion coefficient of Gallium Arsenide (GaAs) more closely matches sapphire than bulk Silicon does. Our chip has twelve thousand transistors in an area of 2.1x2.3 mm and it has 46 perimeter pads and 8 pairs of flip-chip bonding pads in the center for OE array integration. The driver and receiver circuits are 120x350 μm and 60x240 μm respectively. The TX and RX digital core logic is 210x610 μm and 424x710 μm respectively.

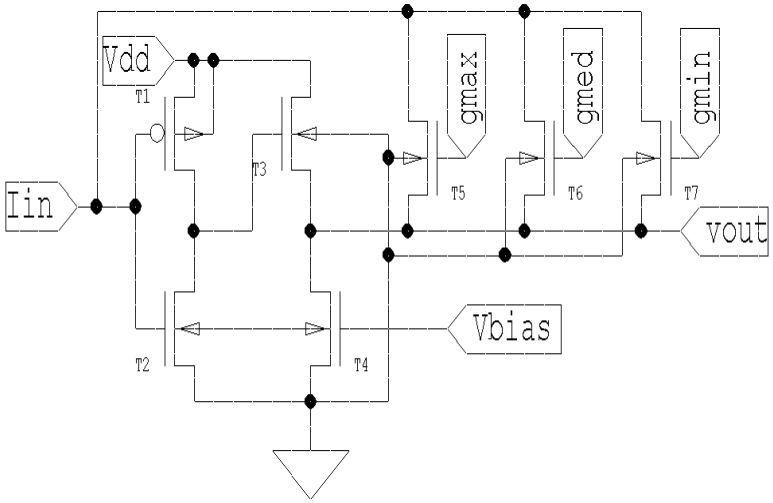


Figure 6-10. Transimpedance amplifier

This chip is a typical mixed-signal circuit design where transceivers circuits are manually laid out by using the layout tool Ledit from Tanner. The algorithm core logic is coded using the hardware description language VHDL and synthesized by Synopsys synthesis tools.

The physical design is done using the Tanner automatic placement and route tool. Special care at the layout level has been taken to ensure signal integrity, such as separating power and ground rails for analog and digital circuits, using interleaving scheme and dual-gate structure for better matched differential circuits, adding dummy gates and inserting necessary decoupling capacitors and electrostatic static discharge circuitry [71].

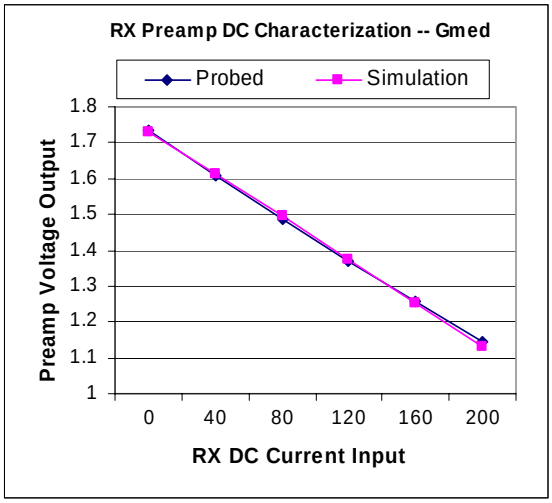


Figure 6-11. Receiver pre-amplifier DC output comparison between simulation and test result

6.D.2. System Integration

6.D.2.a. Integration of CMOS Chip With OE Device

The OE devices used are high-performance 1x4 VCSEL and 1x4 GaAs PIN photodiode arrays with 250 μm pitch from Emcore. The VCSEL wavelength is 850nm and has typical threshold current of 1.5 mA, slope efficiency of 0.45 mW/mA and a differential resistance of 50 ohms. The PIN photodiode has typical responsivity of 0.5 A/W and a capacitance of 0.4 pF. The VCSEL and photodiode array sizes are 1.2x0.45 mm and 1.0x0.45 mm respectively. They are both rated for 3.125 Gbps applications.

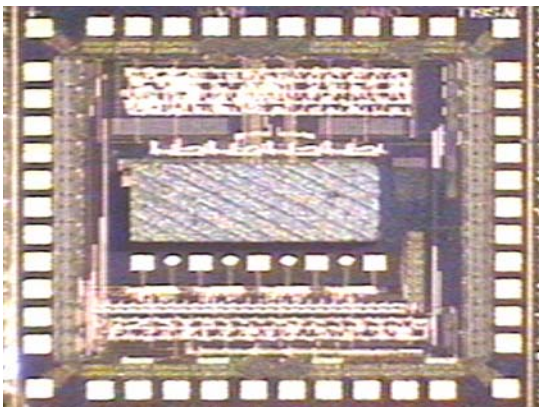
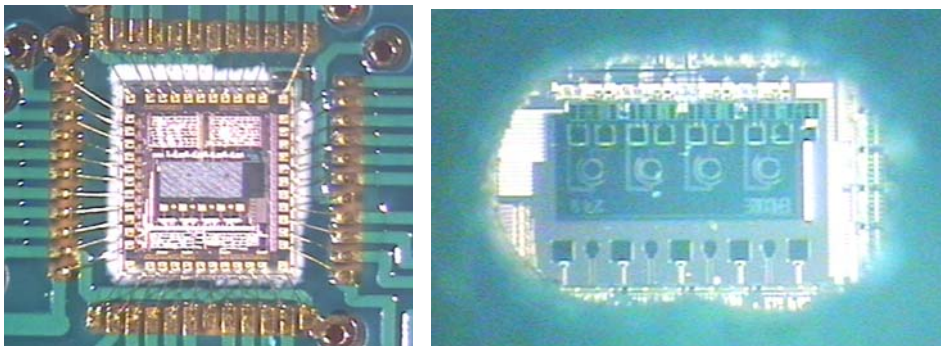


Figure 6-12. CMOS chip with OE device attached

Since flip-chip bonding offers low parasitic capacitance and high integration density, it was used for the integration of CMOS chip with OE devices, which was done by Peregrine semiconductor. Figure 6-12 shows the CMOS chip with the photodetector array attached. Since both VCSEL and photodetector arrays are top emitting devices, light will travel through the substrate of the CMOS chip during their operation.

6.D.2.b. Chip-On-Board (COB) Packaging

A 2x1.5 inch carrier printed circuit board (PCB) was designed and fabricated. The hybrid CMOS chip with OE array attached is placed in the center of carrier PCB and electrical connections between them are made by wire bonding. Directly underneath the hybrid CMOS chip, a 1.6x0.6 mm opening has been made on the PCB to allow the laser beam pass through. Figure 6-13 (a) shows the front view of the CMOS hybrid chip attached onto a carrier PCB with wire bonding. Figure 6-13 (b) shows the photodetector array attached onto CMOS chip by looking through the PCB opening from the back of PCB. As shown in this picture, individual detectors and flip-chip pads can be clearly seen through the sapphire substrate.



(a) (b)

Figure 6-13. Chip-on-board (a) Front view (b) Back view

6.D.2.c. Motherboard

An 8-layer, 7.5x8.5 inch, FR4 motherboard has been designed and fabricated for final system integration. Four carrier PCBs can be perpendicularly connected to the motherboard simultaneously by using Mictor impedance-controlled connectors as shown in Figure 6-14. L-shaped metal clamps are used to make the attachment rigid. A high-performance FPGA (XCV1000E) from Xilinx is used as a center control unit. Connections have been made between the FPGA and each carrier PCB. Various mechanical structures were incorporated to enhance system stability and facilitate system testing. The whole setup was assembled on a breadboard, which provides excellent vibration isolation.

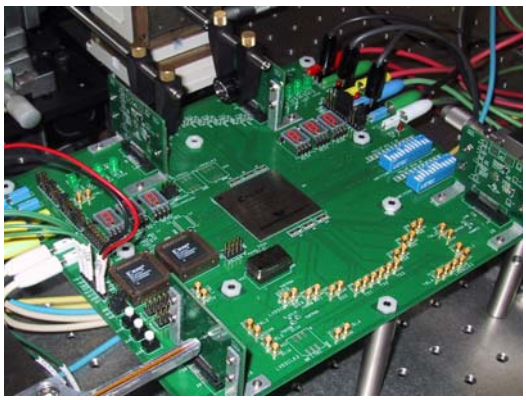
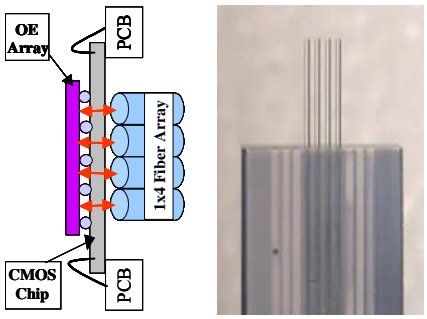


Figure 6-14. Mother board and system setup overview

6.D.3. Fiber Optic Interconnection

Fiber ribbons directly butt-coupled with OE arrays have been widely used in optical parallel module manufacturing with the advancing of packaging technologies. We present here a simplified light-coupling scheme using customized pigtail fiber ribbons. Figure. 6-15 (a) shows a schematic of the light coupling setup where 1x4 fiber array is butt-coupled to the substrate of CMOS chip through an opening of PCB. Given the fact that the VCSEL has a 14°-divergence half-angle and a 200 μm thick sapphire substrate, it is a challenge to couple the light into the fiber efficiently for this arrangement. Although lensed or tapered fiber can improve the light coupling efficiency significantly, they are normally only available for single-mode fiber, which has small core diameter that would require close proximity to VCSEL surface. Therefore, fiber arrays with large core diameter such as multi-mode fiber array are preferred. Another challenge is to maintain good alignment and close proximity between the VCSELs and the fiber ribbon, which is complicated by the fact that the PCB is thick and the opening in the PCB is small.

Therefore, a custom-made fiber optic assembly was carefully selected for this experiment using a 1x4 multi-mode fiber ribbon with 250 μm pitch between fibers. The diameters of the fiber core and cladding are 62.5 μm and 125 μm respectively. A silicon V-groove chip has been used to hold the bare fiber array to ensure the accurate alignment of fiber ribbons. In order for the fiber end to go through PCB opening and reach the CMOS substrate, the bare fiber had to be extended out of the V-groove chip around 3 mm as shown in Figure 6-15 (b). A clean vertical cleaving has been done for the endface of the bare fibers to minimize the optical reflection.



(a) (b)

Figure 6-15. (a) Fiber ribbon light coupling schematic and (b) Pigtail fiber optic assembly

Finally, the fiber assembly was placed and fixed on the top of a metal support that was mounted on top of a micro-positioning stage providing sub-micron fiber positioning with 5 degrees of freedom. A camera was set up to facilitate the alignment process and an active alignment scheme was adopted in order to fine-tune the position of the fiber array. For the transmitter, an optical power meter was used to measure the optical power coupled into the fiber. For the receiver, the voltage output of the pre-amplifier was monitored and the alignment process determines where the maximum voltage drop is observed. Experimental results show that as much as 60% of the light is coupled into the fiber in this arrangement. Therefore, such a setup provides a simple, efficient way to establish the short-range parallel optical links in a laboratory environment.

6.D.4. Free-Space Optical Interconnection

The option of using free-space optical interconnection has also been explored to establish the communication between two chips. Two 7-element Universe Kogaku f-1.2 lenses are used to link a pair of carrier PCBs, separated by 76 mm. A charge coupled device (CCD) camera is used to capture the images of both the laser beams and the OE array to facilitate the alignment process. The voltage drop of the receiver pre-amplifier output was monitored during the active alignment process when the VCSELs are turned on. Due to the small geometry of the laser beams (250 μm pitch for 1x4 array), and the paraxial nature of the optical link, low skew propagation was achieved across the four channels.

6.E. Experimental Results

Extensive testing was done to verify the performance of the transceivers, the free-space and fiber optical interconnects and the power negotiation algorithm. A burn-in test, a test of the power optimization algorithm, and an illustration of the adaptability of the system will be described in this section.

To show reliable, extended operation, all VCSELs in four channels were turned on and driven with pseudo-random data while all receivers, set to medium gain, monitored incoming data and counted errors. The system clock for this test was 100 MHz. The I_{on} and I_b of the VCSELs were set to 4.8 mA and 1.6 mA respectively. The system ran continuously for 15 hours and no errors were observed. Figure 6-16 shows an eye diagram from the optical output of one of the data channels.

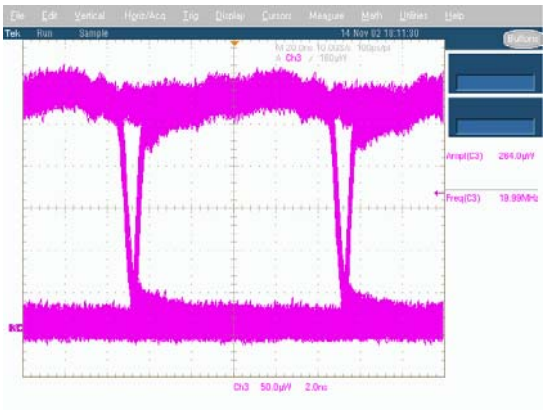


Figure 6-16. Measured eye diagram at 100 MHz

The data rate of test system corresponds to an operating point $B\tau<1.2$ and thus we expect that the algorithm will produce a solution that that is not power optimal. To determine how well the algorithm worked in this non-ideal case, and to verify the power negotiation algorithm, control logic for the decision algorithms were implemented in an FPGA with a fiber optic interconnection used between transceivers. A BER of 10^{-7} was selected as the target error rate and each iteration was run long enough to allow more than 256 errors to be received. The maximum gain setting was selected on the receiver side. The results of each iteration of the algorithm were tracked and monitored using displays on the motherboard. The result of this test was that the algorithm converged to a power setting with an I_{on} of 1.8 mA and an I_b of 0.6 mA that just enables the link to achieve the target BER at a system clock rate of 100 MHz. Another test was run to show that a minimum I_{on} of 2.0 mA can achieve the same BER at zero biasing. This confirms that even for $B\tau<1.2$, the algorithm can produce a nearly optimal power setting. While more tests with higher speed data are desirable to test whether the algorithm will produce an optimal power solution for $B\tau>1.2$, current testing is limited by the speed of standard digital cells in the CMOS chip. The algorithm converges to the same setting each time, which demonstrates the stability of the algorithm implementation with a fixed operating environment. The low modulation current needed can be explained by the low-loss (short) optical path and high gain used in the receiver.

For the free-space setup, an I_{on} of 1.26 mA and an I_b of 1.0 mA were also found. To demonstrate the ability of the power negotiation algorithm to find the optimum power setting when the system operating environment has been changed, an optical filter (attenuator) with an optical density (OD) of 0.2 (equal to 63% transmission rate) was inserted between two compound lenses to decrease the incident optical power on the detector. As shown in Table 6-1, the optimum I_{on} was found at 1.6 mA for the high loss optical path instead of 1.26 mA for low loss path, which illustrated the functionality of the power negotiation algorithm under dynamic conditions.

Table 6-1. Comparison of optimum power settings between with and without optical filter inserted

	Receiver Gain Setting	Optimum Power Setting	
		Modulation Current (mA)	Bias current (mA)
Without Filter	Gmax	0.26	1.0
With Filter	Gmax	0.40	1.2

6.F. Power Optimization Conclusions

A packaging friendly parallel optical transceiver design based on VCSELs and photodetectors has been presented. The transceiver has a built-in power negotiation algorithm based on bit error rate, which can be performed to find the optimum power setting for VCSELs in a dynamic system-operating environment. The theory of optical communication channel has been examined and the dependency of link BER on optical power has been described to show the theoretical background of this power negotiation algorithm. This design, as did the design in the last section, demonstrated advanced methods of improving data communication with reduced power operation, and is therefore directly applicable to the PCA problem domain. By using the techniques and implementation described here, it is possible to build simpler, more efficient optical links by virtue of the ability to save power whenever possible and yet have the flexibility to account for dynamic changes and less than state-of-the-art optical packages.

7. RAW EMULATION BOARD

Tiled processor architectures are important to polymorphous architectures because they allow processor performance to scale as silicon feature sizes continue to shrink. Current microprocessor architectures, which typically use instruction-level parallelism, will not continue to allow the scaling of clock speeds as feature sizes continue to shrink because of the global signal lines required for architectures based on single, sequential instruction streams. Tiled processor architectures allow software to explicitly identify parallelism, which allows smaller independent processors to scale with clock speed. Tiled processor architectures also significantly increase memory bandwidth between the processors and on-board memory. This increased memory bandwidth is particularly important for data-intensive signal-processing applications as computational power increases. MIT's Raw architecture is an example of a tiled architecture.

The Raw emulation board was developed to support an FPGA-based emulation of the Raw chip. The Raw FPGA emulation was used to validate logical functionality of the Raw chip, validate the performance model of the Raw simulator, to support faster benchmarking, and to debug the design of the Raw board before the Raw chip was available. The architecture of the Raw emulator board was developed jointly with the MIT Raw project. A block diagram of the Raw emulator system is shown in Figure 7-1. The Raw emulator is implemented in an IKOS (a company later acquired by Mentor Graphics) FPGA-based system that connects to the Raw emulation board through ribbon cable connectors. The board interfaces with a Linux-based host computer through a an FPGA board. The Raw emulation board has connectors for user interfaces, expansion, and standard main memory modules.

The Raw emulation board was fabricated and is shown in Figure 7-2. The board was successfully demonstrated and led to the development of the Raw Handheld board under the Abstract Machines for Polymorphous computing (AMP) project. Functionality of the Raw emulation board and the Handheld board have been shown to be identical (except for

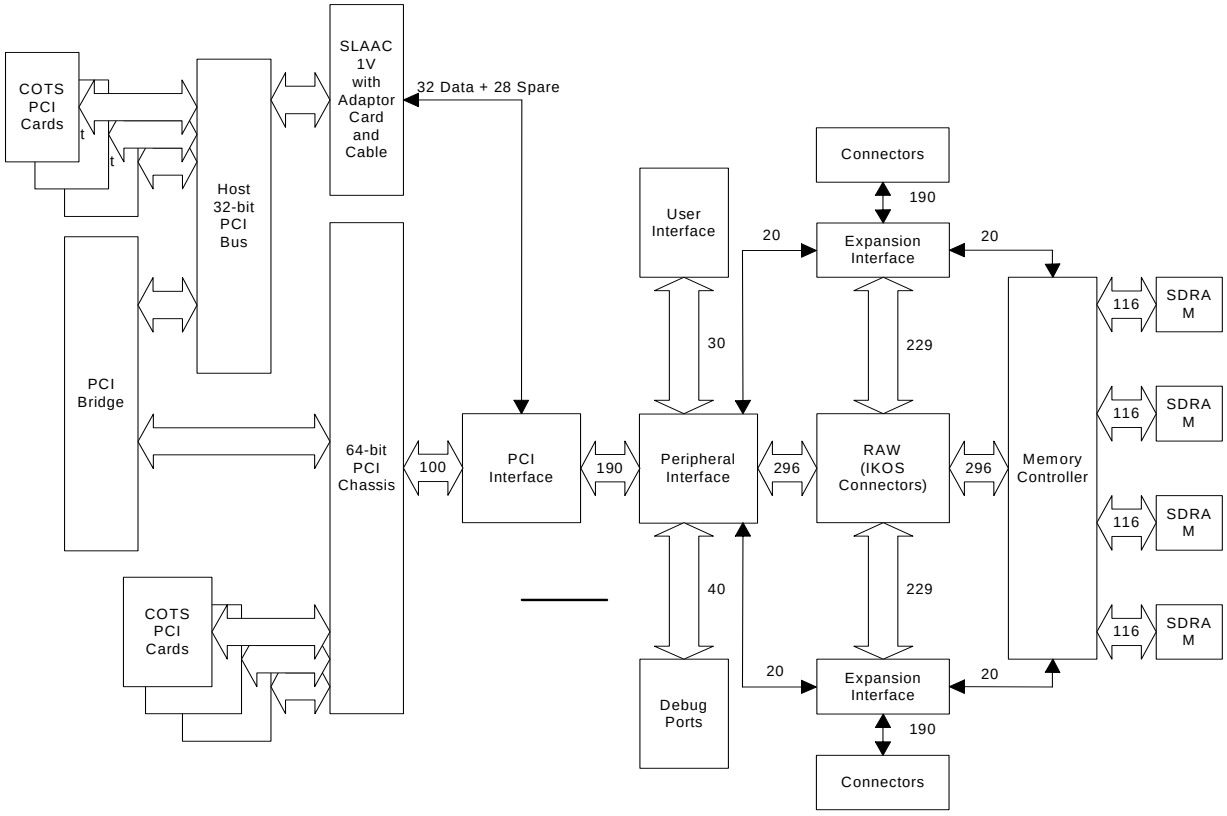


Figure 7-1. Raw emulation system block diagram

performance since the emulator clock speed is slow). More importantly, the development of the emulation system allowed the validation of the logical design of the Raw chip, which later was shown to be functional at first silicon.

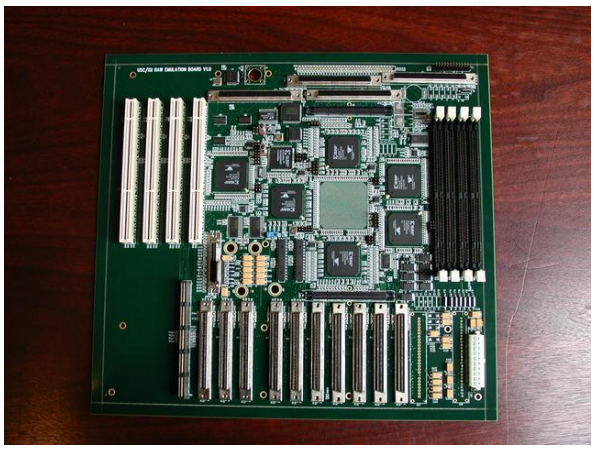


Figure 7-2. Raw emulation board

8. SUMMARY AND CONCLUSIONS

The RATS project focused on designing a system to demonstrate a tile-based architecture and two critical aspects needed to validate the use of ultra-high-capacity global free-space interconnection fabrics for PCA-like architectures. The Raw emulation system led to the validation and debugging of the Raw architecture. The Raw chip and HandHeld board were developed under another project, and the success of that development comes directly from the emulation board developed on the RATS project.

Free-space interconnect issues fell into the categories of interface issues and internal switch issues. It was recognized early on that both issues must be addressed in order to fully exploit the advantages of global optical free-space interconnections. Simply put, having a revolutionary compact Tbps-class switch fabric does no good unless the interface bottlenecks are overcome. Consequently, it turned out that a significant portion of our effort centered on the issues associated with "feeding the switch."

The major research highlights, accomplishments, and implications of the accomplishments in the area of free-space interconnects that have been detailed in this report are:

- Development of a 10 Gbps Optical Network Interface Card (NIC) demonstrating direct leveraging of photonic interconnection fabrics for multi-board systems. This concept allows the extension of the free-space global interconnection fabric to distributed computing applications. For example in a PCA-like application, in which computing resources are distributed across different boards or chassis, the new NIC concept will allow the massive free-space switch core to be used as a reconfiguration engine that rapidly re-allocates resources by reconfiguring the interconnection fabric.
- Development of a multi-scale optical lens system that removes distortion for a free-space optical interconnection module. The validation of this concept showed that a relatively simple, but novel hybrid lens design could achieve the required registration needed to implement thousands of optical links in a global fabric, with reasonable packaging costs. This global fabric would enable global, dynamic communication patterns in a PCA system.
- Demonstration of a free-space optical interconnection assembly with misalignment sensitivities better than 25 microns and 1 degree. The multi-scale optical approach was proven to be highly tolerant to misalignments owing to its effective trade-off of spatial and angular positioning for VCSEL-array-based system. This optical interconnection assembly is a necessary component of the free-space interconnect that would enable global communication in a PCA system that has not previously been demonstrated.
- Integration of 4 1x4 VCSEL arrays and 4 1x4 photodetector arrays on a single Ultra-Thin Silicon (UTSi) circuit for an aggregate free-space I/O bandwidth of 40 Gbps per OEIC. This achievement proved that dense ultra-high bandwidth optical I/O can be integrated with PCA processor components and scaled to densities that suggest that a path to Tbps per chip area I/O will be achievable.
- Extension of a chip-on-pin concept for optoelectronic device packaging for fiber modules and full ~10 cm diameter multi-chip array on a PCB. This achievement validated the system packaging approach that will permit mechanical placement of the smart pixel arrays, using modified conventional packaging techniques that achieve the alignment tolerances necessary for massive global optical interconnects.
- Demonstration of a 160 Gpbs free-space interconnected module. This prototype is the first to show the integration of arrays of high bandwidth (2.5 Gbps per channel) arrays interconnecting chips in a fully connected global fabric. This achievement proved that the basic concept can be implemented and suggests that scaling to Tbps-class fabrics is achievable. These fabrics would enable communication bandwidths far higher than provided today in electrically interconnected PCA systems, and the global connections would support important operations like corner turns and the dynamic allocation of resources that would be required to support full-scale system-level morphing.

Having validated the critical base issues of interfacing and general packaging and alignment for the RATS free-space fabric architecture concept, future work will be needed to focus on the issues of device and system integration, reliability, and power management of very dense photonic arrays. It is envisioned that future requirements will push aggregate chip I/O level into the multi-Tbps regime, where conventional metal-based electrical strategies are severely limited. Specifically, the main question that needs to be answered is: How dense can we make the chip photonic I/O and still maintain reliability and packaging efficiency? Consequently, future work will be concerned with developing integration and packaging that address thermal management, power management protocols, and low cost system packaging and alignment concepts that will permit the full exploitation of photonic fabrics in overcoming communications-bound performance limitations in future computing systems.

REFERENCES

[1] Zuerndorfer, B., G. Shaw, "SAR Processing for RASSP Application," *Proceedings 1st Annual RASSP Conference*, Arlington, VA, August 1994, pp 253-268.

[2] Usui, M., Sato, N., Ohki, A., Matsuura, N., Tanaka, N., Enbutsu, K., Amano, M., Hikita, M., Kagawa, T., Katsura, K., Ando, Y., "ParaBIT-1: 60-Gb/s-Throughput Parallel Optical Interconnect Module," *Proceedings 50th Electronic Components & Technology Conference*, 2000, pp 1252–1258.

[3] Eichenberger, J., Toyoda, S., Takezawa, N., Keller, C., Sugiyama, M., Iwasaki, Y., "Multi-Channel Optical Interconnection Modules Up To 2.4 Gb/s/ch," *Proceedings 51st Electronic Components and Technology Conference*, 2001, pp 880–885.

[4] Eichenberger, J., Takezawa, N., Keller, C., Toyoda, S., Sugiyama, M., "1.25-3.2 Gb/s/ch Multi-Channel Optical Interconnection Modules and Their Mass-Production," *Lasers and Electro-Optics Society Annual Meeting*, 13th Annual Meeting. IEEE, Vol. 2*,* 2000, pp 675 –676.

[5] Ekman, J., Chandramani, P., Gui, P., Wang, X., Kiamilev, F., Christensen, M., Haney, M., Milojkovic, P., Driscoll, K., Vanvoorst, B., Liu, Y., Nohava, J., Cox, J.A., "Gigabit Switch Using Free-Space and Parallel Optical Data Links for a PCI-Based Workstation Cluster," *Lasers and Electro-Optics Society Annual Meeting*, 13th Annual Meeting. IEEE, Vol. 2*,* 2000, pp 494 -495.

[6] Szymanski, T. H., Au, A., ré-Roula, M. Lafrenie, Tyan, V., Supmonchai, B., Wong, J., Zerrouk, B., Obenaus, S. T., "Terabit Optical Local Area Networks for Multiprocessing Systems," *Applied Optics*, Vol. 37, 1998, pp 264–275.

[7] Nishimura, S., Kudoh, T., Nishi, H.; Yamamoto, J., Ueno, R.; Harasawa, K., Fukuda, S., Shikichi, Y., Akutsu, S., Tasho, K., Amano, H., "RHiNET-3/SW: an 80-Gbit/s High-Speed Network Switch for Distributed Parallel Computing," *Hot Interconnects,* Vol. 9, 2001, pp 119-123.

[8] *The Asyncronous Transfer Mode (ATM) Standard,* EIA Standard.

[9] *The Gigabit Ethernet Standard,* IEEE Standard 802.32, 1998.

[10] *The High Performance Parallel Interface (HIPPI) Standard,* ANSI Standard.

[11] Lehman, J., "An Introduction to the ChEEtah Project," *Hot Interconnects Symposium V*, August 1997.

[12] Liu, Y., Lehman, J., Cox, A., Hibbs-Brenner, M., "Opto-Electronic Component, Integration and Packaging Technology Advancements, and Their Impact on Massively Parallel Interconnects," *Parallel Interconnects*, 1999, pp 3-4.

[13] *The PCI Bus Standard,* IEEE Standard P1386.1.

[14] The 10GbE LR Pro Server Adapter Specifications, Intel Inc., [Online] http://www.intel.com/network/connectivity/products/10gigabit/index.htm.

[15] The OC-192/STM-64 Interface Card, CIENA Corporation, [Online]. http://www.ciena.com/products/switching/metrodirectork2/oc192card.

[16] Raghavan, Y.G., Kim, Chuang, T.Y., Madhavan, B., Levi, A. F. J., "A Gbyte/s Parallel Fiber-Optic Network Interface for Multimedia Applications," *IEEE Network,* 1999, pp 13, 20-28.

[17] Au, Albert, Supmonchai, Boonchuay, Szymanski, Ted H., "Testbed for a Scalable Terabit Optical Local Area Network," *Applied Optics*, Vol. 39, No. 23, 2000, pp 4131-4142.

[18] "The PCI Express Specifications," The PCI Special Interests Group, [Online] http://www.pcisig.com.

[19] *The VHDL Standard*, IEEE Standard 1076, 1993.

[20] *The JTAG Standard,* IEEE Standard 1149.1, 1990.

[21] Honeywell International Inc. (1997, MN), "VCSEL-Based Interconnects in VLSI Architectures for Computational Enhancement (VIVACE)," VLSI Photonics Program, [Online] http://www.htc.honeywell.com/photonics/vlsi.html.

[22] Haney, Michael W., Christensen, Marc P., Milojkovic, Predrag, Fokken, Gregg J., Vickberg, Mark, Gilbert, Barry K., Rieve, James, Ekman, Jeremy, Chandramani, Premanand, Kiamilev, Fouad, "Description and Evaluation of the *FAST-Net* Smart Pixel-Based Optical Interconnection Prototype," *Proceedings of the IEEE,* Vol. 88, No. 6, June 2000, pp 819-828.

[23] *The Infiniband Standard*, The Infiniband Trade Association, [Online] http://www.infinibandta.com/.

[24] *The PCI-X 2.0 Standard*, The PCI Special Interests Group, [Online] http://www.pcisig.com/.

[25] "The IXA Network Processor Series," Intel Inc (2002, Santa Clara, CA), [Online] http://www.intel.com/design/network/products/npfamily/.

[26] *The Virtex-II Pro Data Book*, Xilinx Inc, 2002.

[27] Nakahara, T., Matsuo, S., Fukushima, S., Kurokawa, T., "Performance Comparison Between Multiple-Quantum-Well Modulator-Based and Vertical-Cavity-Surface-Emitting Laser-Based Smart Pixels," *Applied Optics*, Vol. 35, 1996, pp 860-871.

[28] Haney, M. W., Christensen, M. P., "Performance Scaling Comparison for Free-Space Optical and Electrical Interconnection Approaches," *Applied Optics*, Vol. 37, 1998, pp 2886–2894.
[29] Haney, M. W., Christensen, M. P., Milojkovic, P., Ekman, J., Chandramani, P., Rozier, R., Kiamilev, F., Liu, Y., Hibbs-Brenner, M., "Multi-Chip Free-Space Global Optical Interconnection Demonstration with Integrated Arrays of Vertical-Cavity Surface-Emitting Lasers and Photodetectors," *Applied Optics*, Vol. 38, 1999, pp 6190-6200.
[30] Haney, M. W., Christensen, M. P., Milojkovic, P., Fokken, G. J., Vickberg, M., Gilbert, B. K., Rieve, J., Ekman, J., Chandramani, P., Kiamilev, F., "Description and Evaluation of the FAST-Net Smart Pixel-Based Optical Interconnection Prototype," *Proceedings of the IEEE*, Vol. 88, pp 819-828.
[31] Christensen, M. P., Milojkovic, P. M., Haney, M. W., "Low-Distortion Hybrid Optical Shuffle Concept," *Optics Letters*, Vol. 24, 1999, pp 169-171.
[32] Christensen, M. P., Milojkovic, P., Haney, M. W., "Analysis of a Hybrid Micro/Macro-Optical Method for Distortion Removal in Free-Space Optical Interconnections," *JOSA*, No. 12. December 2002.
[33] Christensen, M. P., McFadden M. J., Haney, M. W., "Experimental Validation of a Hybrid Micro/Macro-Optical Concept for Minimizing Distortion in the *FAST-Net* Global Interconnection System," *Proceedings Optics in Computing*, Lake Tahoe, Nevada, January 2001.
[34] M. P. Christensen, M. J. McFadden, P. Milojkovic, M. W. Haney, "Experimental Validation of a Hybrid Micro/Macro-Optical Method for Distortion Removal in Free-Space Optical Interconnections," *Applied Optics*, No. 35, December 2002.
[35] Milojkovic, P., Christensen, M. P., Haney, M. W., "Minimum Lens Complexity Design Approach for a Free-Space Macro-Optical Multi-Chip Global Interconnection Module," *Proceedings Optics in Computing*, Quebec, Canada, June 2000, pp 917-938.
[36] Milojkovic, P., Ph.D. Dissertation, George Mason University, August 2001.
[37] Andreou A.G., Kalayjian Z.K., Apsel A., Pouliquen P.O., Athale R.A., Simonis G., Reedy R., "Silicon On Sapphire CMOS for Opto-Electronic Microsystems," *IEEE Circuits & Systems Magazine*, Vol. 1, No. 3, 2001, pp 22-30.
[38] "HyperTransport Technical White Paper," [Online] www.hypertransport.org.
[39] "RapidIO Technical White Paper," [Online] www.rapidio.org.
[40] POS-PHY Level 4 Resource Center, [Online] www.pmc-sierra.com/posphylevel4.
[41] Rambus Yellowstone Interface, [Online] http://www.rambus.com.
[42] Nordin, R.A., Levi, A. F. J., Nottenburg, R. N., O'Gorman, J., Tanbun-Ek, T., Logan, R.A., "A System Perspective on Digital Interconnection Technology," *IEEE Journal of Lightwave Technology*, Vol. 10, 1992, pp 811-827.
[43] Govindarajan, M., Siala, S., Nottenburg, R N., "Optical Receiver Systems for High Speed Parallel Digital Data Links," *IEEE Journal of Lightwave Technology*, Vol. 13, No. 7, July 1995, pp 1555-1565.
[44] Karstensen, H., Hanke, C., Honsberg, M., Kropp, J.R., Wieland, J., Blaser, M., Weger, P., Popp, J., "Parallel Optical Interconnection for Uncoded Data Transmitted with 1 Gb/s-per-Cannel Capacity, High Dynamic Range, and Low Power," *Journal of Lightwave* Technology, Vol. 13, No.6, June 1995, pp 1017-1030.
[45] Nishimura, Shinji, Inoue, Hiroaki, Matsuoka, Hiroshi, Yokota, Takashi, "Synchronized Parallel Optical Interconnection Subsystem Implemented in the RWC-1 Massively Parallel Computer," *IEEE Photonics Technology Letter*, Vol. 11, No. 10, October 1999, pp 360-367.
[46] Nishimura, Shinji, Inoue, Hiroaki, Matsuoka, Hiroshi, Yokota, Takashi, "Optical Interconnection Subsystem Used in the RWC-1 Massively Parallel Computer," *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 5, No. 2, March/April 1999, pp 360-367.
[47] Kuznia, "Ultra-Thin Silicon-on-Sapphire (UTSi) CMOS," *CO-OP/Peregrine/USC Workshop*, University of Southern California, June 2001.
[48] McGettigan, "Eight Channel, One Clock, One Frame LVDS Transmitter/Receiver," [Online] xilinx.com/apps/XPP245.
[49] Herzen, B.V., Brunetti, J., "Multi-Channel 622 Mb.s LVDS Data Transfer for Virtex–E Devices," [Online] xilinx.com/apps/XPP233.
[50] Kiamilev, E., Krishnamoorthy, A. V., "A High-Speed 32-Channel CMOS VCSEL Driver with Built-In Self-Test and Clock Generation Circuitry," *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 5, No. 2, March/April 1999, pp 287-295.
[51] Plant, D. V., Venditti, M. B., Laprise, E., Faucher, J., Razavi, K., Chateauneuf, M., Kirk, A. G., Ahearn, J. S., "256-Channel Bi-Directional Optical Interconnect Using VCSELs and Photodiodes on CMOS," *Journal of Lightwave Technology*, Vol. 19, No.8, August 2001, pp 1093–1103.
[52] Temkin, Wilmsen H., Colden, L.A., *Vertical-Cavity Surface-Emitting Lasers: Design, Fabrication, Characterization, and Applications,* Cambridge University Press, 1999.
[53] "1x4 VCSEL Array Data Sheet," EMCORE Corporation, Somerset, NJ, USA, [Online] www.emcore.com.

[54] "1x4 PIN Photodetector Array Data Sheet," EMCORE Corporation, Somerset, NJ, USA, [Online] www.emcore.com.
[55] Krishnamoorthy, A.V., Chirovsky, L.M.F., Hobson, W.S., Leibenguth, R.E., Hui, S.P., Zydzik, G.J., Goossen, K.W., Wynn, J.D., Tseng, B.J., Lopata, J., Walker, J.A., Cunningham, J.E., D' Asaro, L. A., "Vertical-Cavity Surface-Emitting Lasers Flip-Chip Bonded to Gigabit-per-Second CMOS Circuits," *IEEE Photonics Technology Letters*, Vol. 11, No. 1, January 1999, pp 128-130.
[56] Meindl, J. D. et al, "Interconnecting Device Opportunities for Gigascale Integration (GSI)," *International Electron Devices Meeting*, Session 23, Washington, DC, December 2001, pp 525–528.
[57] Esener, S. C., "Implementation and Prospects for Chip-to-Chip Free-Space Optical Interconnects," *International Electron Devices Meeting*, Session 23, Washington, DC, December 2001, pp 541–544.
[58] Venditti, M. B., "Design and Verification of an OE-VLSI Chip with 1080 VCSELs and PDs Heterogeneously Integrated with CMOS," *The 14th Annual Meeting of The IEEE Lasers & Electro-Optics Society*, post deadline session, San Diego, CA, November 2001.
[59] Haney, M. W. et al, "Opto-Mechanical Design and Implementation of the FAST-Net Smart Pixel-Based Free-Space Optical Interconnection Prototype," *Proceedings of IPACK'01*, Kauai, Hawaii, July 2001.
[60] Ekman, J. et al, "System Design and Packaging for an Optically Interconnected MCM Switch for Parallel Computing," *Proceedings of IPACK'01*, Kauai, Hawaii, July, 2001.
[61] Ishii, Y., Koike, S., Arai, Y., Ando, Y., "SMT-Compatible Optical-I/O Chip Packaging for Chip-Level Optical Interconnects," *Electronic Components and Technology Conference*, 2001.
[62] Haake, J. Neranek, M., "In Package Micro-Aligner for Fiber-Optic Packaging," *Electronic Components and Technology Conference*, 1998.
[63] "Modulating VCSELs," internal Honeywell document.
[64] Liu, Y. S., Wojnarowshi, R. J., Hennerssy, W.A., Piacentr, P.A., "Plastic VCSEL Array Packaging and High Density Polymer Waveguides for Board and Backplane Optical Interconnect," *Electronic Components and Technology Conference*, 1998.
[65] Chua, L., Fork, D. K., Hantschel, T., "Densely Packed Optoelectronic Interconnect Using Micromachined Springs," *IEEE Photonics Technology Letters*, Vol. 14, No. 6, June 2002, pp 846–848.
[66] Verdiell, J., Webjorn, J., Kohler, R., Epitaux, M., Finot, M., Kirkpatrick, P., Lake, R., Colin, S., Mader, T., Bennett, J., "Automated Opto-Electronic Packaging for 10Gb/s Applications," *Electronic Components and Technology Conference*, 2001.
[67] Agrawal, G. P., *Fiber-optic Communication Systems*, John Wiley & Sons, Inc., New York, 1992.
[68] Obermann, K., Kindt, S., Petermann, K., "Turn-On Jitter in Zero-Biased Single-Mode Semiconductor Lasers," *IEEE Photonics Technology Letters*, Vol. 8, No. 1, January 1996 pp 31-33.
[69] Chen, L. P., Lau, K. Y., "Regime Where Zero-Bias Is the Low-Power Solution for Digitally Modulated Laser Diodes," *IEEE Photonics Technology Letters*, Vol. 8, No. 2, February 1996, pp 185-187.
[70] Dutta, N. K., "Power Penalty Due to Timing Jitter for Laser Modulated Without Prebias," *Applied Physics Letters* Vol. 67, No. 22, November 1995.
[71] Hastings, *The Art of Analog Layout*, Prentice Hall, 2000.

APPENDIX I: ACRONYMS

2D-FFT	Two-dimensional Fast Fourier Transform
AMCC	Applied Micro Circuits Corporation
ASIC	Application Specific Integrated Circuit
ATM	Asynchronous Transfer Mode
BER	Bit Error Rate
BERT	Bit Error Rate Tester
CAD	Command Address and Data
CCD	Charge Coupled Device
CLK	Clock
CML	Current Mode Level
CMOS	Complimentary Metal Oxide Semiconductor
CTL	Control
DAC	Digital to Analog Converter
DARPA	Defense Applied Research Projects Agency
DDR	Double Data Rate
FASTNET	Free-space Accelerator for Switching Terabit NETworks
FIFO	First-In First-Out
FOCUTS	Flipped Opto-electric Chips on Ultra Thin Silicon
FPGA	Field Programmable Gate Array
FSOI	Free-space Optical Interconnections
GaAs	Gallium Arsenide
HIPPI	High Performance Parallel Interface
IC	Integrated Circuit
I/O	Input/Output
JTAG	Joint Test Action Group
LFSR	Linear Feedback Shift Register
MCM	Multi Chip Module
MGT	Multi Gigabit Transceivers
MPI	Message Passing Interface
MT	Mechanically Transferable
NA	Numerical Aperture
NIC	Network Interface Card
OE	Opto-electronic
OEIC	Opto-electric Integrated Circuit
ONIC	Vertical Network Interface Card
OPL	Optical Path Length
PCA	Polymorphous Computer Architectures
PCB	Printed Circuit Board
PCI	Peripheral Component Interconnect
PD	Photo-detector
PISO	Parallel In, Serial Out
PLL	Phase Locked Loop
PRBS	Pseudo Random Bit Stream
PROM	Programmable Read-Only Memory
RATS	Reactive ArchiTectureS
RC	Resistor-Capacitor
RMS	Root Mean Square
SERDES	SERializer/DESerializer
SDR	Single Data Rate
SOS	Silicon on Sapphire
SPA	Smart Pixel Array
TIA	Transimpedance Amplifier

UTSi	Ultra-Thin Silicon
VCSEL	Vertical Cavity Surface Emitting Laser
VHDL	Very high-speed integrated circuits Hardware Description Language
VIVACE	Vertical Cavity Surface Emitting Lasers (VCSEL)-based Interconnects in VLSI Architectures for Computational Enhancement
VLSI	Very Large Scale Integrated circuits
VSR	Very Short Reach